Pattern recognition and machine learning for medical image processing

Daniel RACOCEANU

Professor,
Sorbonne University

DEEP LEARNING

Update Oct. 2020

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

0

References used in this course

- Jürgen Schmidhuber's blog
  - http://people.idsia.ch/~juergen/
  - Scientific Director of the Swiss AI Lab IDSIA, Lugano, Switzerland
  - Jürgen Schmidhuber (2015). Deep learning in neural networks: An overview, Review Article, Neural Network, Volume 61, pp. 85-117
  - PyBrain - a modular Machine Learning Library for Python
    - PyBrain is short for Python-Based Reinforcement Learning, Artificial Intelligence and Neural Network Library.

PyBrain

1

References used in this course

- Yann LeCun
  - PhD @ Univ. Pierre and Marie Curie, Paris
  - Professor at the New York University
  - http://yann.lecun.com/
  - Director of AI Research, Facebook

- Xiaogang Wang
  - Chinese University of Hong Kong
  - http://www.ee.cuhk.edu.hk/~xgwang/

2

References

- D. E. Rumelhart, G. E. Hinton, R. J. Williams, "Learning Representations by Back-propagation Errors," Nature, Vol. 323, pp. 533-536, 1986.
- K. Fukushima, "Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position," Biological Cybernetics, Vol. 36, pp. 193-202, 1980.
- N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. Piater, A. J. Rodriguez-Sanchez, L. Wiskott, "Deep Hierarchies in the Primate Visual Cortex: What Can We Learn For Computer Vision?" IEEE Trans. PAMI, Vol. 35, pp. 1847-1871, 2013.
- A. Krizhevsky, L. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Proc. NIPS, 2012.
- Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based Learning Applied to Document Recognition," Proceedings of the IEEE, Vol. 86, pp. 2278-2324, 1998.
- G. E. Hinton, S. Osindero, and Y. Teh, "A Fast Learning Algorithm for Deep Belief Nets," Neural Computation, Vol. 18, pp. 1527-1544, 2006.

3

Outline

- Introduction to deep learning
- Deep learning for object recognition
- Deep learning for object segmentation
- Deep learning for object detection
- Open questions and future works

4

Outline

- Introduction to deep learning
  - Historical review of deep learning
  - Introduction to classical deep models
  - Why does deep learning work?
- Deep learning for object recognition
- Deep learning for object segmentation
- Deep learning for object detection
- Open questions and future works

5

Outline

- Introduction to deep learning
  - Historical review of deep learning
  - Introduction to classical deep models
  - Why does deep learning work?
- Deep learning for object recognition
- Deep learning for object segmentation
- Deep learning for object detection
- Open questions and future works

6

Machine Learning

$x \longrightarrow \boxed{F(x)} \longrightarrow y$

y: Class label (Classification); Vector (Estimation)

Object recognition → {dog, cat, horse, flower, …}

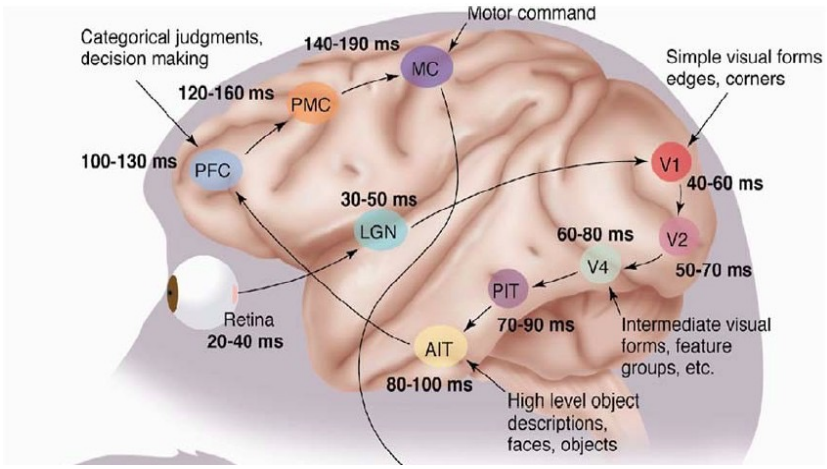
Low-resolution image

Super resolution →

High-resolution image

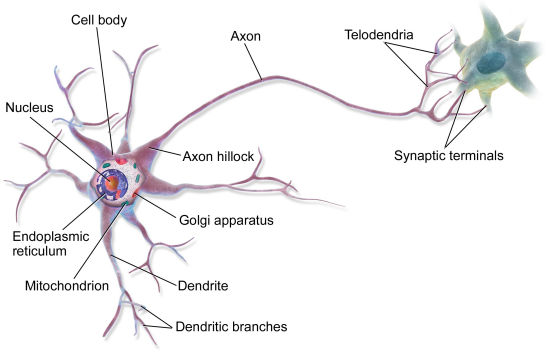
7

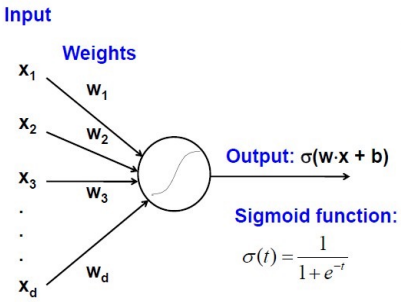
Inspiration (Loosely): The Brain



8

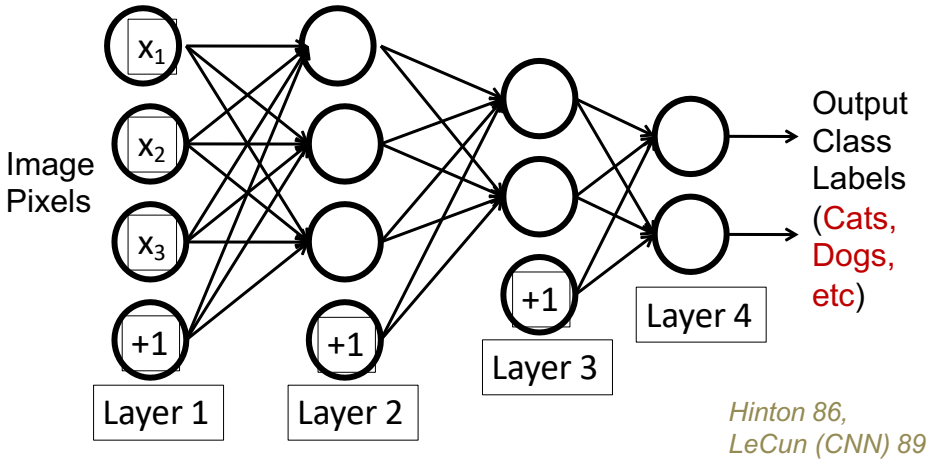
Modeling the neuron (perceptron)





9

4-layer Fully Connected Neural Net



Hinton 86, LeCun (CNN) 89

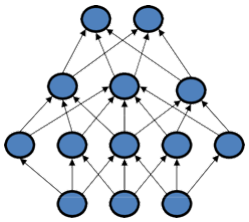
10

Geoffrey Hinton - U.Toronto & Google

Neural network Back propagation

Nature

198 6



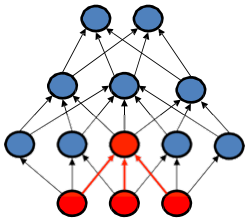
- Solve general learning problems
- Tied with biological system

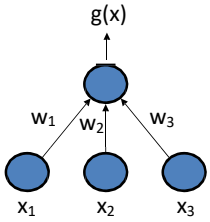
12

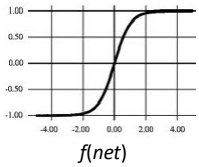
Neural network
Back propagation

Nature

1986

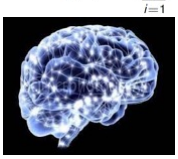


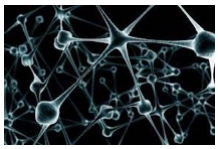




f(*net*)

$$g(\mathbf{x}) = f(\sum_{i=1}^{d} x_i w_i + w_0) = f(\mathbf{w}^t \mathbf{x})$$



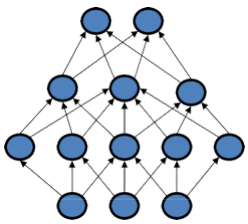


13

Neural network
Back propagation

Nature

1986



- Solve general learning problems
- Tied with biological system

But it is given up…

- Hard to train
- Insufficient computational resources
- Small training sets
- Does not work well

14

Was it too early ? ….

- Concorde
  - Concorde speed was beyond that of twilight, and may equal or exceed the speed of rotation of the earth. On flights to the west, it was possible to arrive at a local time earlier than the departure time. **On certain transatlantic flights departing from Heathrow or Paris, you could take off even after sunset and still catch the sunset, by landing in the day**. This was well publicized by British Airways, which used the slogan **"Arrive before you leave."**

concorde sonic boom

15

Was it too early ? ….

- Minitel in France … 70' – precursory of www. … ?

16

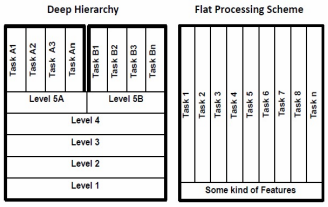
Neural network
Back propagation

Nature

1986 2006

- SVM
- Boosting
- Decision tree
- KNN
- …

- Flat structures
- Loose tie with biological systems
- Specific methods for specific tasks
  - Hand crafted features (GMM-HMM, SIFT, LBP, HOG)*



Kruger et al. TPAMI'13

*GMM = Gaussian Mixture Model
HMM = Hidden Markov Model
SIFT = Scale-invariant feature transform
LBP = Local binary patterns
HOG = Histogram of oriented gradients

17

Yann LeCun - Pierre-and-Marie-Curie University -> New York University
Facebook Artificial Intelligence

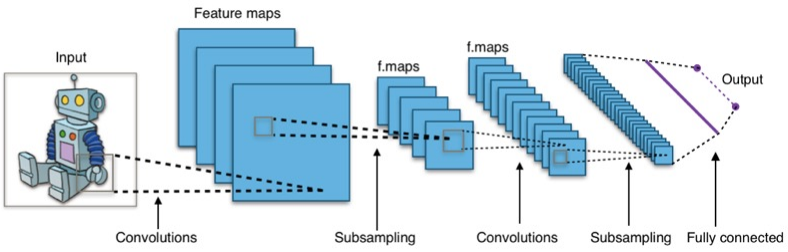
Neural network
Back propagation

Nature

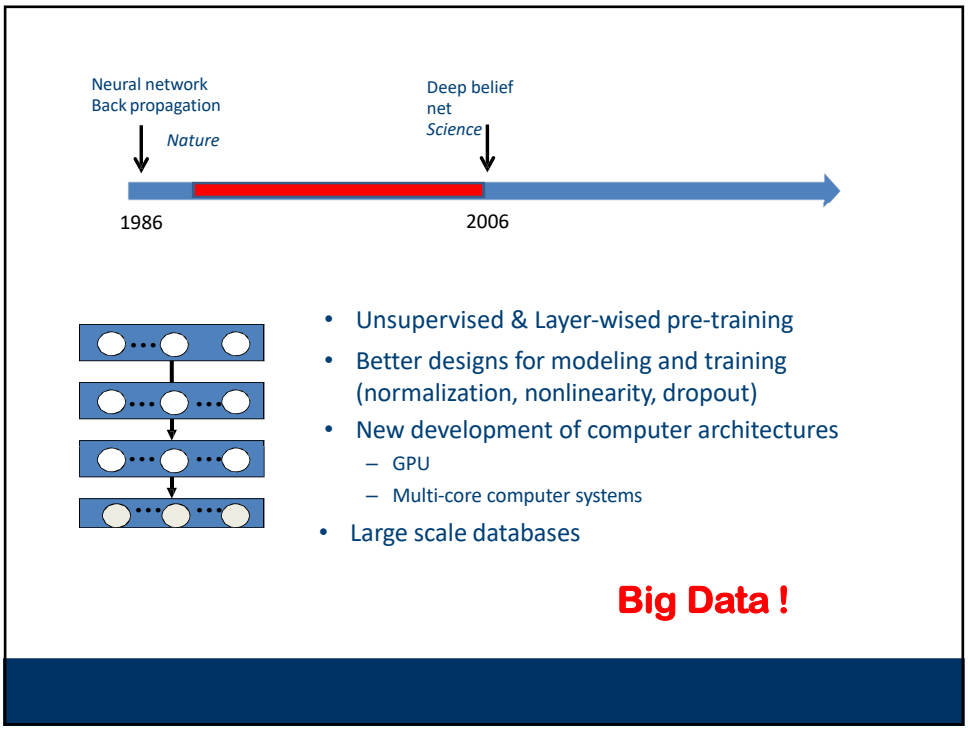
Convolutional
Neural Nets

1986 1989

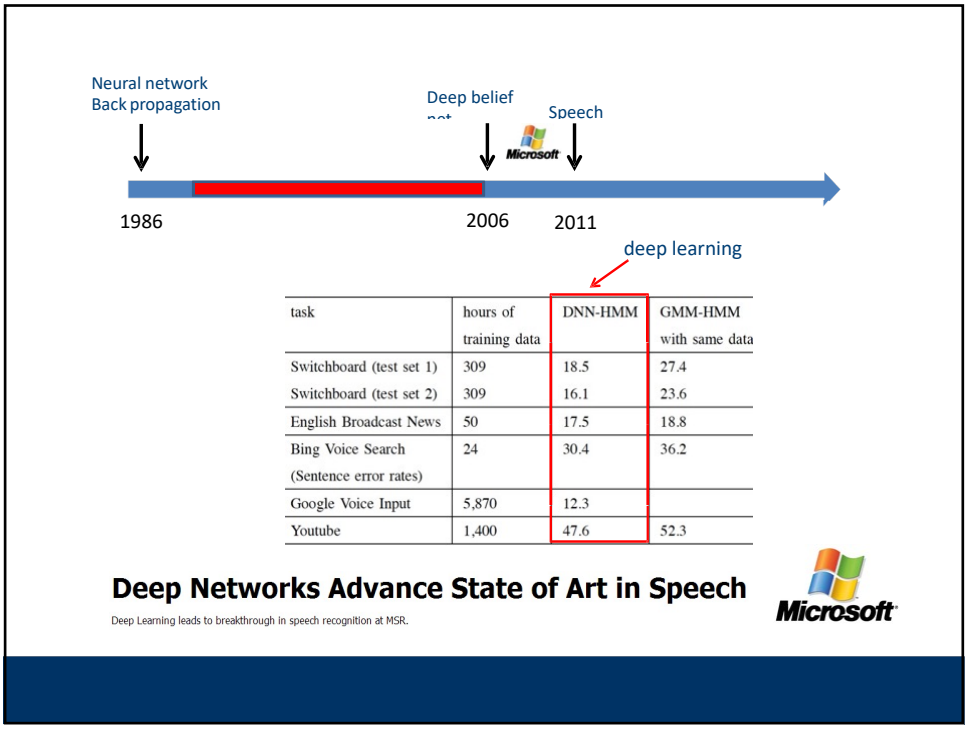
- A typical CNN architecture
  - Convolutional, Pooling, ReLU (Rectified Linear Units),
  - Fully connected, Loss layer (Softmax, Sigmoid cross-entropy, Euclidian)



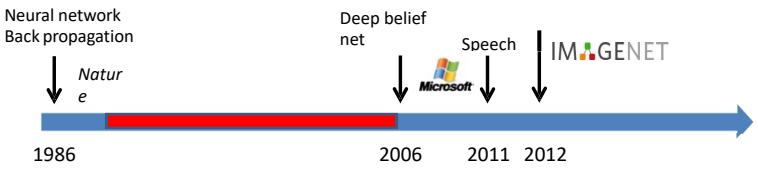
18


Neural network
Back propagation
Nature
Deep belief
net
Science
1986
2006
Unsupervised & Layer-wised pre-training
Better designs for modeling and training (normalization, nonlinearity, dropout)
New development of computer architectures
GPU
Multi-core computer systems
Large scale databases
Big Data !

19


Neural network
Back propagation
Deep belief
Speech
Microsoft
1986
2006
2011
deep learning
task
hours of training data
DNN-HMM
GMM-HMM with same data
Switchboard (test set 1) 309 18.5 27.4
Switchboard (test set 2) 309 16.1 23.6
English Broadcast News 50 17.5 18.8
Bing Voice Search (Sentence error rates) 24 30.4 36.2
Google Voice Input 5,870 12.3
Youtube 1,400 47.6 52.3
Deep Networks Advance State of Art in Speech
Deep Learning leads to breakthrough in speech recognition at MSR.
Microsoft

20



Rank	Name	Error rate	Description
1	**U. Toronto**	**0.15315**	**Deep learning**
2	U. Tokyo	0.26172	Hand-crafted features and learning models. Bottleneck.
3	U. Oxford	0.26979	
4	Xerox/INRIA	0.27058	

Object recognition over 1,000,000 images and 1,000 categories (2 GPU)

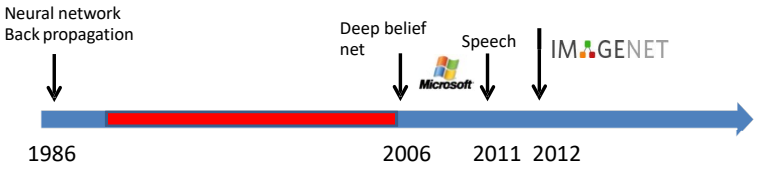
A. Krizhevsky, L. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," NIPS*, 2012

*NIPS = Neural Information Processing Systems.

21

Examples from ImageNet

22



- ImageNet 2013 – image classification challenge

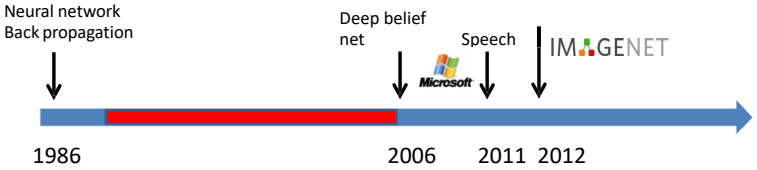
Rank	Name Description	Error rate	
1	NYU	0.11197	Deep learning
2	NUS	0.12535	Deep learning
3	Oxford	0.13555	Deep learning

MSRA, IBM, Adobe, NEC, Clarifai, Berkley, U. Tokyo, UCLA, UIUC, Toronto …. Top 20 groups all used deep learning

- ImageNet 2013 – object detection challenge

Rank	Name	Mean AP	Description
1	UvA-Euvision	0.22581	Hand-crafted features
2	NEC-MU	0.20895	Hand-crafted features
3	NYU	0.19400	Deep learning

23



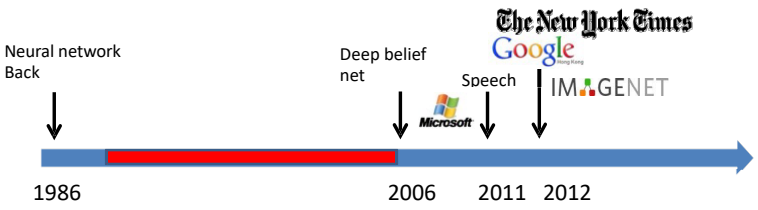
- ImageNet 2014 – Image classification challenge

Rank	Name Description	Error rate	
1	Google	0.06656	Deep learning
2	Oxford	0.07325	Deep learning
3	MSRA	0.08062	Deep learning

- ImageNet 2014 – object detection challenge

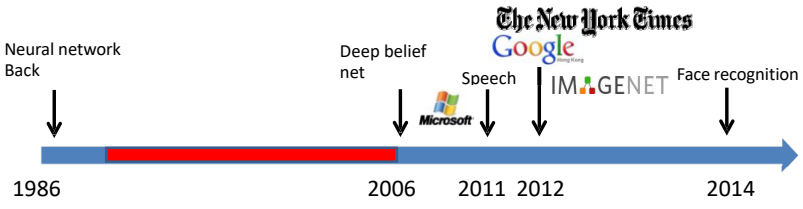
Rank	Name	Mean AP	Description
1	Google	0.43933	Deep learning
2	CUHK	0.40656	Deep learning
3	DeepInsight	0.40452	Deep learning
4	UvA-Euvision	0.35421	Deep learning
5	Berkley Vision	0.34521	Deep learning

24



- Google and Baidu announced their deep learning based visual search engines (2013)
  - Google
    - "on our test set we saw **double the average precision** when compared to other approaches we had tried. We acquired the rights to the technology and went full speed ahead adapting it to run at large scale on Google's computers. We took cutting edge research straight out of an academic research lab and launched it, in just a little over six months."
  - Baidu

25



- Deep learning achieves 99.15% face verification accuracy on Labeled Faces in the Wild (LFW), close to human performance

  Y. Sun, X. Wang, and X. Tang. Deep Learning Face Representation by Joint Identification-Verification. NIPS, 2014.

26

Labeled Faces in the Wild (2007)

Best results without deep learning

Random guess (50%)

Eigenface (60%)

TL Joint Bayesian (96.33%), 2013
Human cropped (97.53%)

DeepID - Deep Learning (99.15%)
Human funneled (99.20%)

27

Unrestricted, Labeled Outside Data Results

2014

Attribute classifiers[11]	0.8525 ± 0.0060
Simile classifiers[11]	0.8414 ± 0.0041
Attribute and Simile classifiers[11]	0.8554 ± 0.0035
Multiple LE + comp[14]	0.8445 ± 0.0046
Associate-Predict[18]	0.9057 ± 0.0056
Tom-vs-Pete[23]	0.9310 ± 0.0135
Tom-vs-Pete + Attribute[23]	0.9330 ± 0.0128
combined Joint Bayesian[26]	0.9242 ± 0.0108
high-dim LBP[27]	0.9517 ± 0.0113
DFD[33]	0.8402 ± 0.0044
TL Joint Bayesian[34]	0.9633 ± 0.0108
face.com r2011b[19]	0.9130 ± 0.0030
Face++[40]	0.9727 ± 0.0065
DeepFace-ensemble[41]	0.9735 ± 0.0025
ConvNet-RBM[42]	0.9252 ± 0.0038
POOF-gradhist[44]	0.9313 ± 0.0040
POOF-HOG[44]	0.9280 ± 0.0047
FR+FCN[45]	0.9645 ± 0.0025
DeepID[46]	0.9745 ± 0.0026
GaussianFace[47]	0.9852 ± 0.0066
DeepID2[48]	0.9915 ± 0.0013

Table 6: Mean classification accuracy $\hat{u}$ and standard error of the mean S_E.

http://vis-www.cs.umass.edu/lfw/results.html

Unrestricted, Labeled Outside Data Results

2016

Simile classifiers[11]	0.8472 ± 0.0041
Attribute and Simile classifiers[11]	0.8554 ± 0.0035
Multiple LE + comp[14]	0.8445 ± 0.0046
Associate-Predict[18]	0.9057 ± 0.0056
Tom-vs-Pete[23]	0.9310 ± 0.0135
Tom-vs-Pete + Attribute[23]	0.9330 ± 0.0128
combined Joint Bayesian[26]	0.9242 ± 0.0108
high-dim LBP[27]	0.9517 ± 0.0113
DFD[33]	0.8402 ± 0.0044
TL Joint Bayesian[34]	0.9633 ± 0.0108
face.com r2011b[19]	0.9130 ± 0.0030
Face++[40]	0.9950 ± 0.0036
DeepFace-ensemble[41]	0.9735 ± 0.0025
ConvNet-RBM[42]	0.9252 ± 0.0038
POOF-gradhist[44]	0.9313 ± 0.0040
POOF-HOG[44]	0.9280 ± 0.0047
FR+FCN[45]	0.9645 ± 0.0025
DeepID[46]	0.9745 ± 0.0026
GaussianFace[47]	0.9852 ± 0.0066
DeepID2[48]	0.9915 ± 0.0013
TCIT[53]	0.9333 ± 0.0124
DeepID2+[55]	0.9947 ± 0.0012
betaface.com[56]	0.9808 ± 0.0016
DeepID3[57]	0.9953 ± 0.0010
insky.so[59]	0.9551 ± 0.0013
Uni-Ubi[60]	0.9900 ± 0.0032
FaceNet[62]	0.9963 ± 0.0009
Tencent-BestImage[63]	0.9965 ± 0.0025
Baidu[64]	0.9977 ± 0.0006
AuthenMetric[65]	0.9977 ± 0.0009
MMDFR[67]	0.9902 ± 0.0019
CW-DNA-1[70]	0.9950 ± 0.0022
Faceall[71]	0.9940 ± 0.0010
JustMeTalk[72]	0.9887 ± 0.0016
Facevisa[74]	0.9917 ± 0.0019
pose+shape+expression augmentation[75]	0.9807 ± 0.0060
ColorReco[76]	0.9940 ± 0.0022
Asaphus[77]	0.9815 ± 0.0039
Daream[78]	0.9968 ± 0.0009
Dahua-FaceImage[80]	0.9978 ± 0.0007
Easen Electron[81]	0.9968 ± 0.0009

Table 6: Mean classification accuracy $\hat{u}$ and standard error of the mean S_E.

28

Face Recognition (99%)

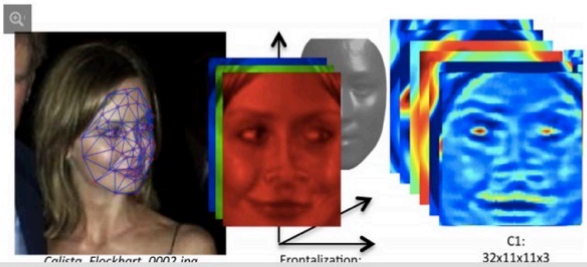
GIZMODO

+ FOLLOW

Facebook's Facial Recognition 'Approaching Human-Level Performance'

Jamie Condliffe
Filed to: FACEBOOK 3/18/14 8:32am

13,028 1 ★



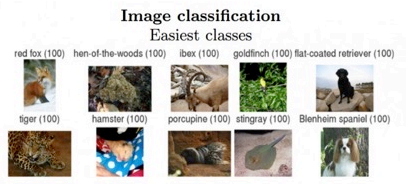
29

Object Recognition (~95%)

MIT Technology Review

The Revolutionary Technique That Quietly Changed Machine Vision Forever

Machines are now almost as good as humans at object recognition, and the turning point occurred in 2012, say computer scientists.



30

House Number Recognition (94.8%)

Goodfellow et al.

ICLR 2014

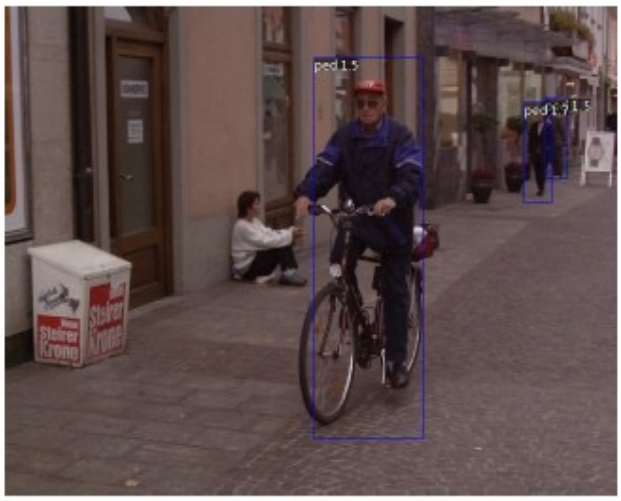
31

Traffic Sign Classification: (97.2%)

Wan et al. 2013

32

Pedestrian Detection (99.x%)



Sarmanet et al. 2013

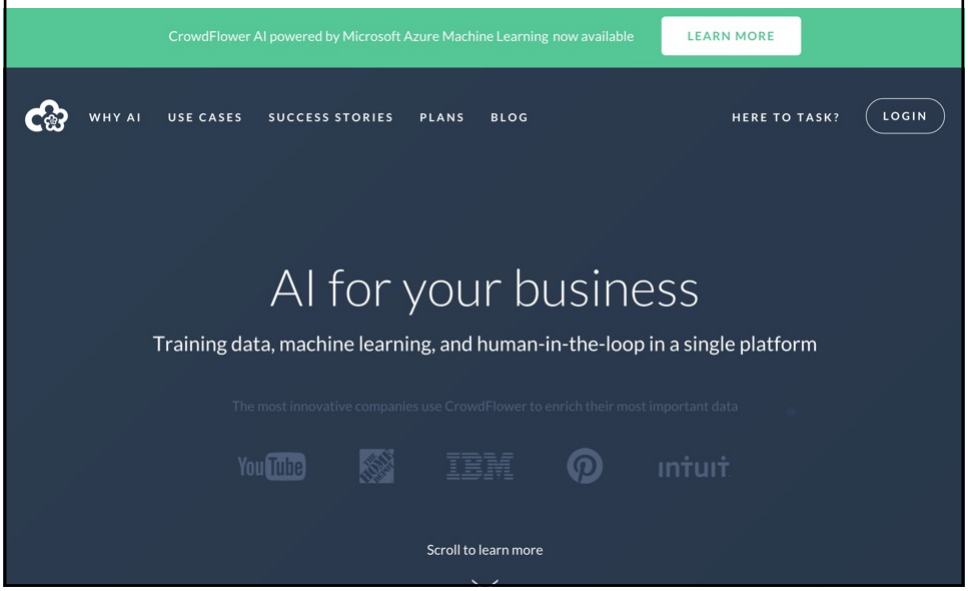
33

Image to Captions (Computer Generated) vs.

34

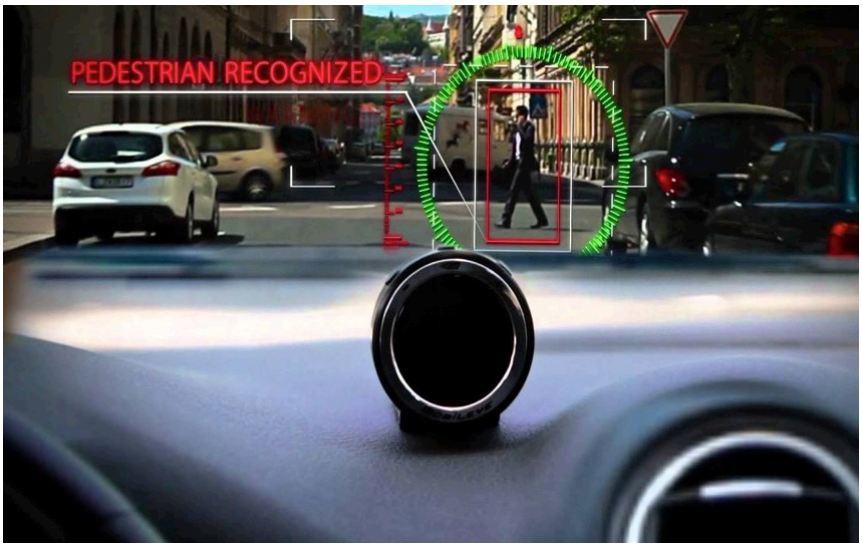
vs. Crowd-sourcing (human) annotation

https://www.crowdflower.com/



35

Hardware: Mobileye



36

MIT Technology Review

BUSINESS NEWS

Is Google Cornering the Market on Deep Learning?

A cutting-edge corner of science is being wooed by Silicon Valley, to the dismay of some academics.

By Antonio Regalado on January 29, 2014

How much are a dozen deep-learning researchers worth? Apparently, more than $400 million.

The acquisition, aimed at adding skilled experts rather than specific products, marks an acceleration in efforts by Google, Facebook, and other Internet firms to monpololize the biggest brains in artificial intelligence research.

38

MIT Technology Review

BUSINESS NEWS

Is Google Cornering the Market on Deep Learning?

A cutting-edge corner of science is being wooed by Silicon Valley, to the dismay of some academics.

By Antonio Regalado on January 29, 2014

How much are a dozen deep-learning researchers worth? Apparently, more than $400 million.

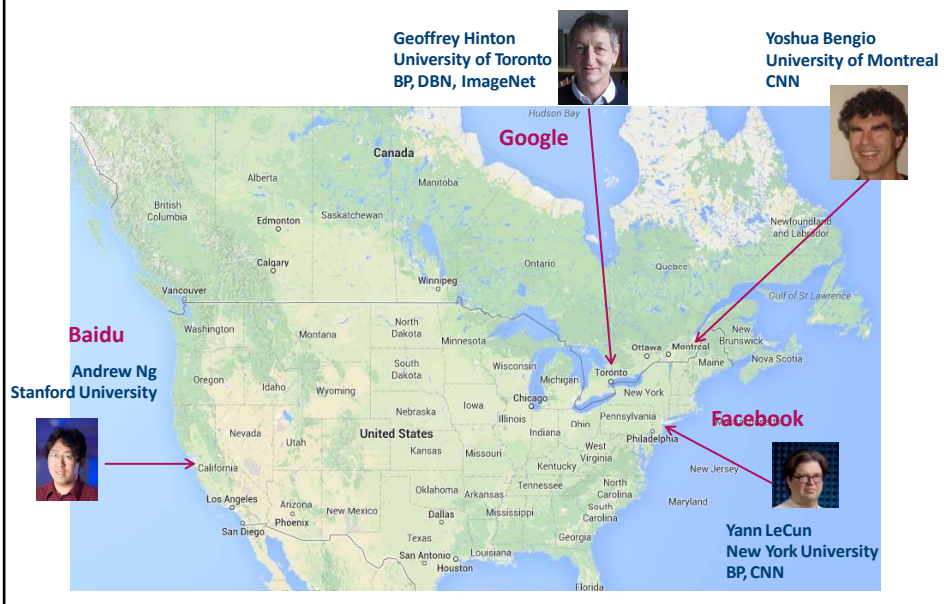
Yoshua Bengio, an AI researcher at the University of Montreal, **estimates that there are only about 50 experts worldwide in deep learning, many of whom are still graduate students**. He estimated that DeepMind employed about a dozen of them on its staff of about 50. "I think this is the main reason that Google bought DeepMind. It has one of the largest concentrations of deep learning experts," Bengio says.

39

News on Deep Learning

Baidu established Institute of Deep Learning	2012
Hinton's group won ImageNet Contest	Oct. 2012
Hinton joined Google	March 2013
Google announced deep learning based visual search engine	March 2013
Baidu announced deep learning based visual search engine	June 2013
Yahoo acquired startup LookFlow working on deep learning	Oct. 2013
Facebook established a new AI lab in NewYork and recruited Yann LeCun	Dec. 2013
Google Acquires DeepMind for USD 400 Million	Jan. 2014
Baidu established a new lab at Shenzhen	2014
Baidu established a new lab at silicon valley and Andrwe Ng is the director	May 2014
Deep learning reached human performance on face verification on LFW	June 2014

40



41

Deep Blue vs. Kasparov - 1997

- 1997: Deep Blue–Kasparov (3½–2½)
- First computer program **to defeat a world champion** in a match under tournament regulations

42

AlphaGo vs. Lee Sedol

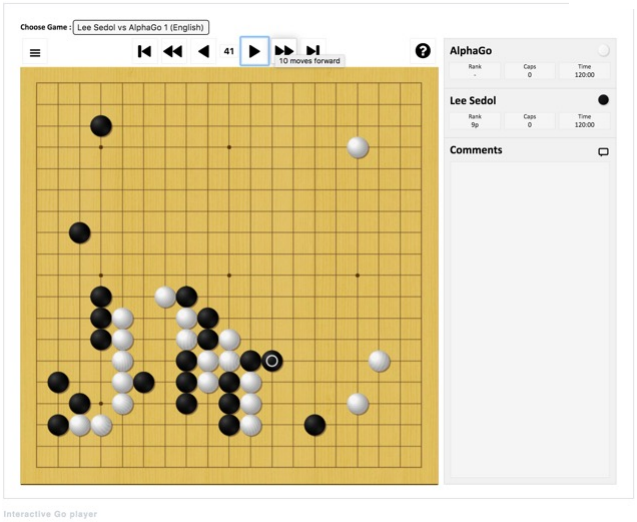
AlphaGo

- Lee Sedol played an historic five game match against Google DeepMind's AlphaGo computer program in March 2016. AlphaGo won the match, making it the **first time a computer Go program had defeated a world class human player** on even terms: **AlphaGo 4 – Lee Sedol 1**

43

AlphaGo vs. Lee Sedol

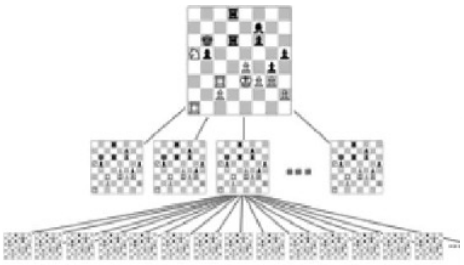
AlphaGo



44

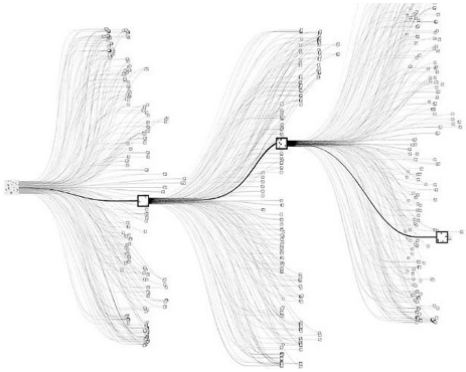
Difference between Deep Blue and AphaGo

- Deep Blue is algorithm- and AlphaGo is Deep Learning-based



Logic-based game …

Tree-based



Intuitive game … Deep Reinforcement Learning

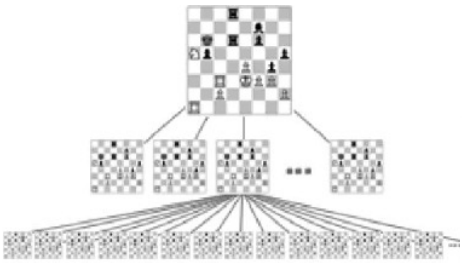
Possible position from each position – 200

Ability to learn and discover strategies by itself

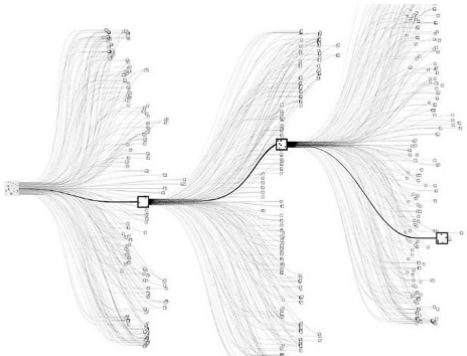
45

Difference between Deep Blue and AphaGo

- Deep Blue is algorithm- and AlphaGo is Deep Learning-based



Possible positions from each position - 20



The number of possible configuration on the board is more than the number of atoms in the universe ! - 10^{81} possible positions

Google's 'AlphaGo' AI trained on 30 million moves

46

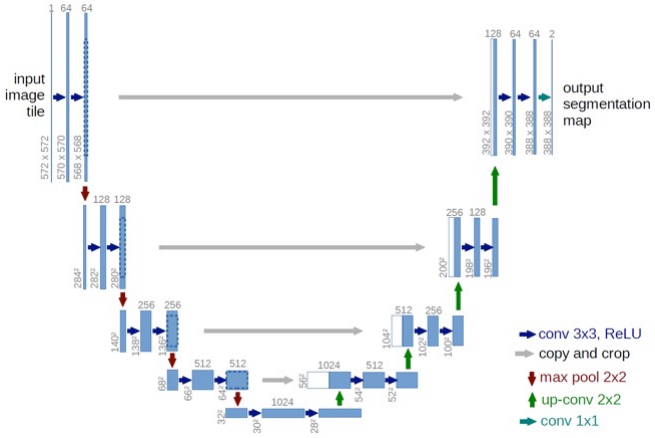
DeepMind Health

- https://deepmind.com/applied/deepmind-health/

Industry leaders create a non-profit organisation that will work to advance public understanding of artificial intelligence technologies (AI) and formulate best practices on the challenges and opportunities within the field.

47

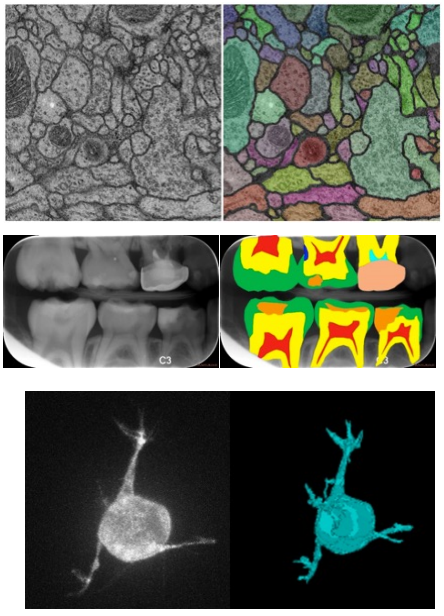
U-Net: Convolutional Networks for Biomedical Image Segmentation



U-Net: Convolutional Networks for Biomedical Image Segmentation
Olaf Ronneberger, Philipp Fischer, Thomas Brox, Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, LNCS, Vol.9351: 234--241, 2015,

48

- U-net - convolutional network architecture for fast and precise segmentation of images.
  - Outperformed the prior best method (sliding-window convolutional network) on the ISBI challenge for segmentation of neuronal structures in electron microscopic stacks.
  - Won - Grand Challenge for Computer-Automated Detection of Caries in Bitewing Radiography ISBI 2015
  - Won - Cell Tracking Challenge ISBI 2015 on the two most challenging transmitted light microscopy categories (Phase contrast and DIC microscopy)



49

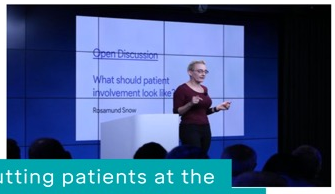
DeepMind Health

- Latest research projects

Announcing DeepMind Health research partnership with Moorfields Eye Hospital

5 July 2016

Applying machine learning to radiotherapy planning for head & neck cancer



Putting patients at the heart of DeepMind Health

21 September 2016

50

Outline

- Introduction to deep learning
  - Historical review of deep learning
  - Introduction to classical deep models
  - Why does deep learning work
- Deep learning for object recognition
- Deep learning for object segmentation
- Deep learning for object detection
- Open questions and future works

52

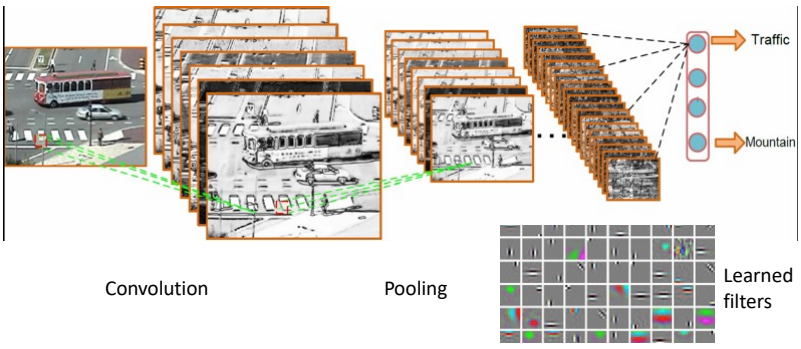
Introduction on Classical Deep Models

- Convolutional Neural Networks (CNN)
  - Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based Learning Applied to Document Recognition," Proceedings of the IEEE, Vol. 86, pp. 2278-2324, 1998.
- Deep Belief Net (DBN)
  - G. E. Hinton, S. Osindero, and Y. Teh, "A Fast Learning Algorithm for Deep Belief Nets", Neural Computation, Vol. 18, pp. 1527-1544, 2006.
- Auto-encoder
  - G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," Science, Vol. 313, pp. 504-507, July 2006.

53

Classical Deep Models

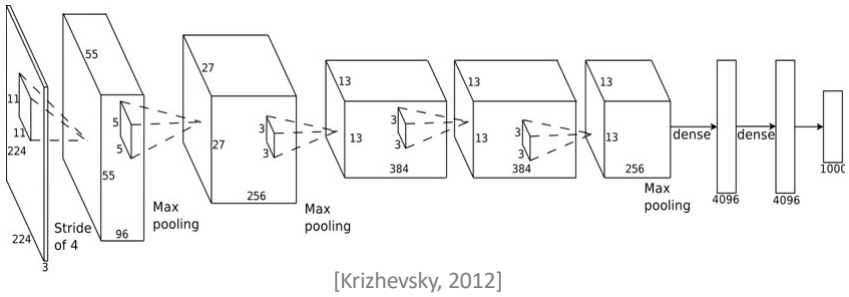
- Convolutional Neural Networks (CNN)
  - First proposed by Fukushima in 1980
  - Improved by LeCun, Bottou, Bengio and Haffner in 1998



54

Typical Deep Network

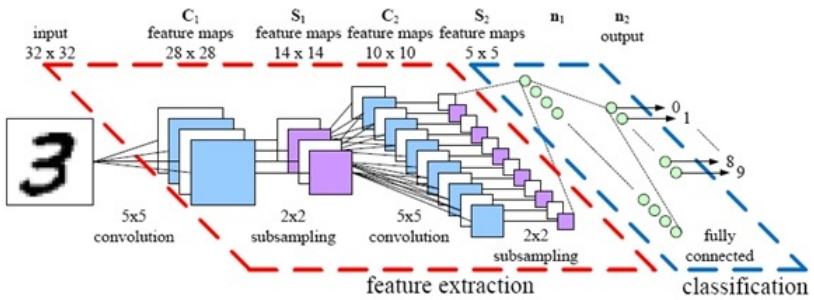
- Convolutional, pooling, dense layers
- Many hidden layers, millions of parameters



[Krizhevsky, 2012]

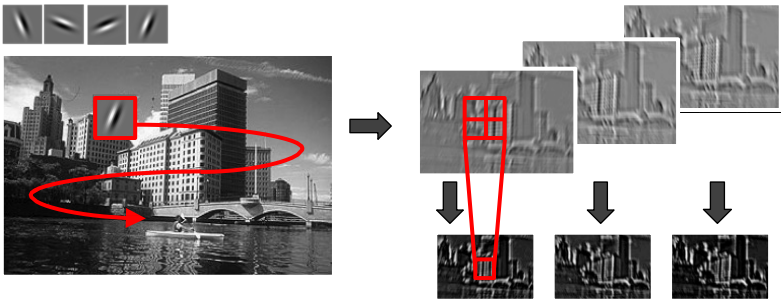
55

Typical Deep Network



56

Convolution & Pooling

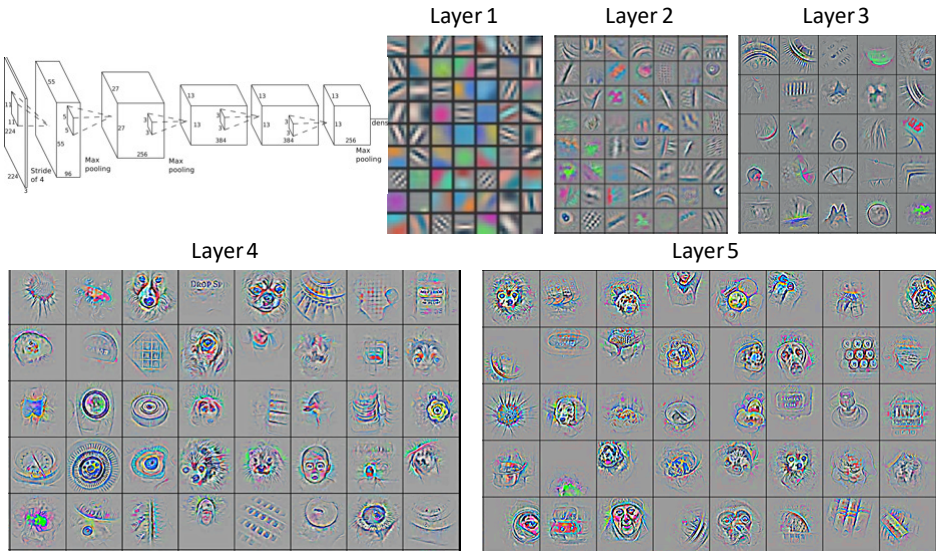


- Convolution
  - Local connectivity
  - Stationarity of the signal
- Pooling
  - Dimensionality reduction
  - Local invariance

57

Learning Abstract Visual Representations

[Zeiler, 2013]

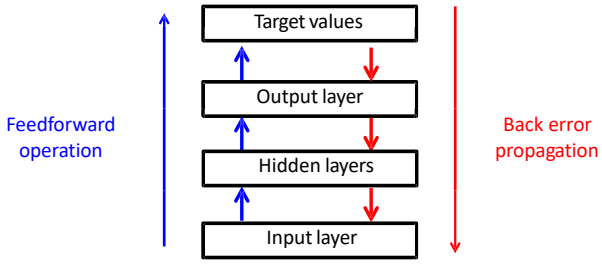


58

Backpropagation

$$\mathbf{W} \leftarrow \mathbf{W} - \eta \bigtriangledown J(\mathbf{W})$$

$\mathbf{W}$ is the parameter of the network; J is the objective function



D. E. Rumelhart, G. E. Hinton, R. J. Williams, "Learning Representations by Back-propagation Errors," Nature, Vol. 323, pp. 533-536, 1986.

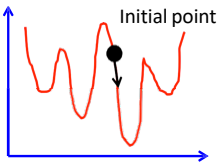
61

Classical Deep Models

- Deep belief net
  - Hinton'06

Pre-training:

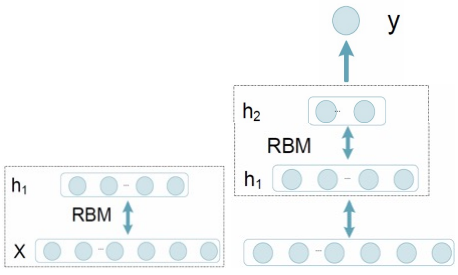
- Good initialization point
- Make use of unlabeled data



$$P(\mathbf{x},\mathbf{h}_1,\mathbf{h}_2) = p(\mathbf{x}|\mathbf{h}_1)\, p(\mathbf{h}_1,\mathbf{h}_2)$$

$$P(\mathbf{x},\mathbf{h}_1) = \frac{e^{-E(\mathbf{x},\mathbf{h}_1)}}{\sum_{\mathbf{x},\mathbf{h}_1} e^{-E(\mathbf{x},\mathbf{h}_1)}}$$

$$E(\mathbf{x},\mathbf{h}_1) = \mathbf{b}'\mathbf{x} + \mathbf{c}'\mathbf{h}_1 + \mathbf{h}_1'\mathbf{W}\mathbf{x}$$



62

Classical Deep Models

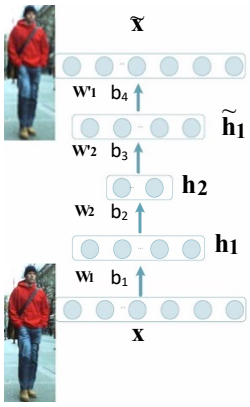
- Auto-encoder
  - Hinton and Salakhutdinov

Encoding: $\mathbf{h}_1 = \sigma(\mathbf{W}_1\mathbf{x}+\mathbf{b}_1)$

$\mathbf{h}_2 = \sigma(\mathbf{W}_2\mathbf{h}_1+\mathbf{b}_2)$

Decoding: $\tilde{\mathbf{h}}_1 = \sigma(\mathbf{W'}_2\mathbf{h}_2+\mathbf{b}_3)$

$\tilde{\mathbf{x}} = \sigma(\mathbf{W'}_1\mathbf{h}_1+\mathbf{b}_4)$



63

Outline

- Introduction to deep learning
  - Historical review of deep learning
  - Introduction to classical deep models
  - Why does deep learning work?
- Deep learning for object recognition
- Deep learning for object segmentation
- Deep learning for object detection
- Open questions and future works

64

Outline

- Introduction to deep learning
  - Historical review of deep learning
  - Introduction to classical deep models
  - Why does deep learning work?
    a. Feature Learning vs. Feature Engineering
    b. Deep Structures vs. Shallow Structures
    c. Joint Learning vs. Separate Learning
    d. High dimensional data transforms
- Deep learning for object recognition
- Deep learning for object segmentation
- Deep learning for object detection
- Open questions and future works

65

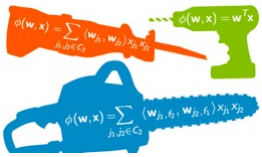
a. Feature Learning vs Feature Engineering

- The performance of a pattern recognition system heavily **depends on feature representations**

66

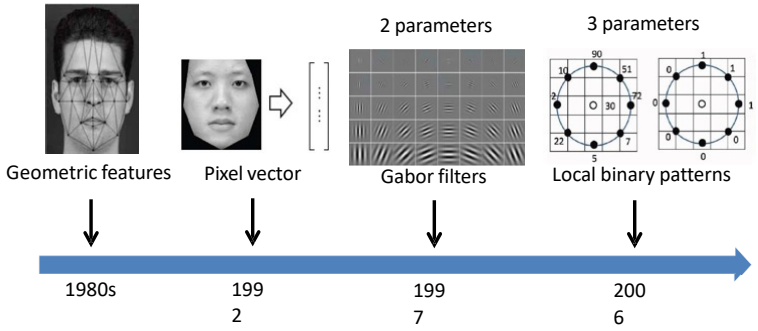
Feature Engineering

- Manually designed features dominate the applications of image and video understanding in the past
  - Reply on human domain knowledge much more than data
  - Feature design is separate from training the classifier
  - If handcrafted features have multiple parameters, it is hard to manually tune them
  - Developing effective features for new applications is slow



67

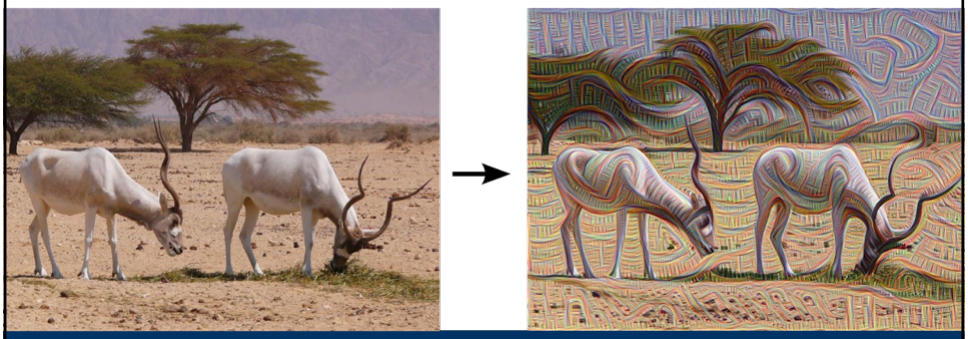
Handcrafted Features for Face Recognition



68

Feature Learning

- Learning transformations of the data that make it easier to extract useful information when building classifiers or predictors
  - Jointly learning feature transformations and classifiers makes their **integration** optimal
  - Learn the values of a **huge number of parameters** in feature representations
  - Faster to get feature representations for **new applications**
  - Make better use of **big data**

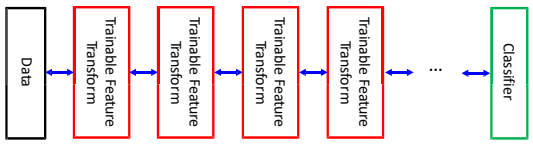


69

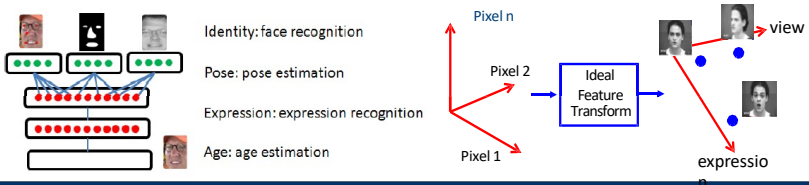
Deep Learning Means Feature Learning

- Deep learning is about learning hierarchical feature representations

$$\mathbf{y} = F(\mathbf{W}^k \cdot F(\mathbf{W}^{k-1} \cdot F(\ldots F(\mathbf{W}^0 \cdot \mathbf{x}))$$

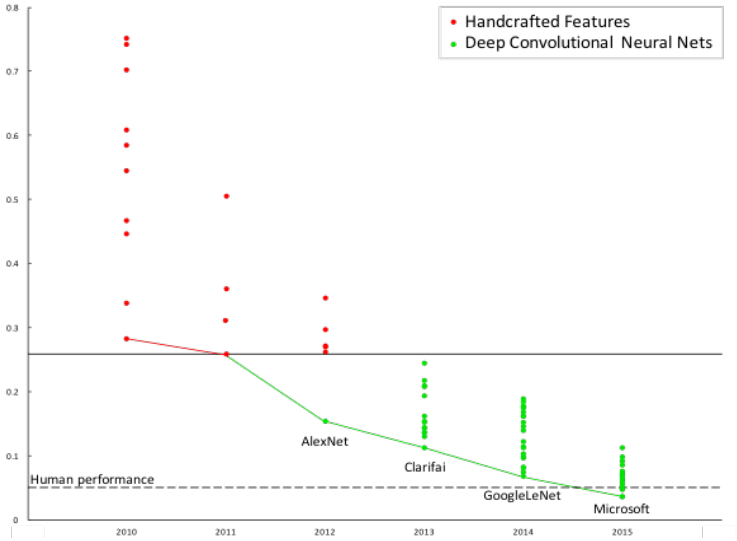


- Good feature representations should be able to disentangle multiple factors coupled in the data



70

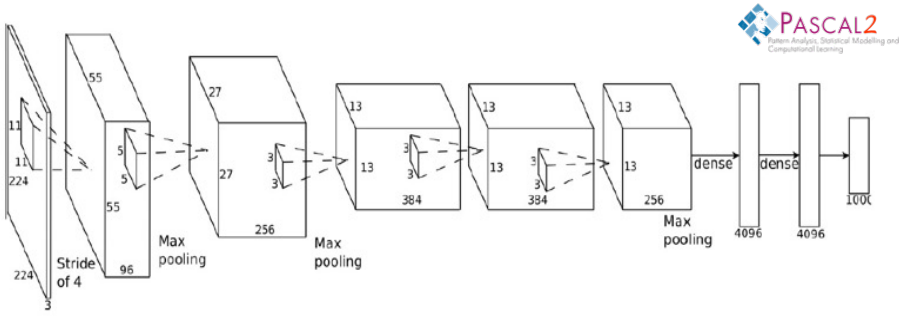
ImageNet Results over the Years



71

Example 1: deep learning generic image features

- Hinton group's groundbreaking work on ImageNet
  - They did not have much experience on general image classification on ImageNet
  - It took one week to train the network with 60 Million parameters
  - The learned feature representations are effective on other datasets (e.g. Pascal Visual Object Classes) and other tasks (object detection, segmentation, tracking, image retrieval)

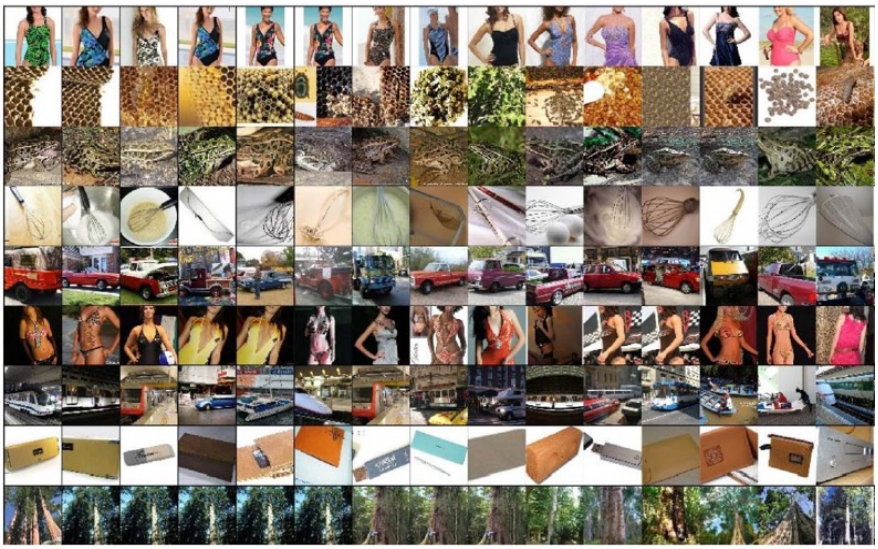


72

96 learned low-level filters

73

Top hidden layer can be used as feature for retrieval



74

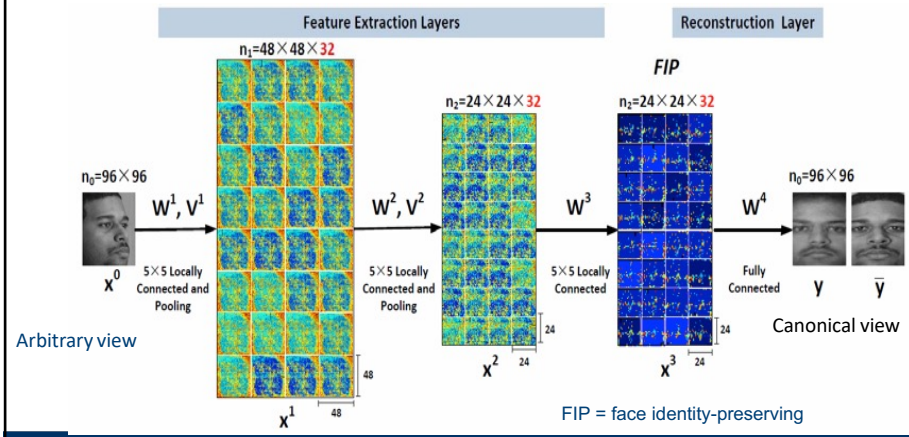
Example 2: deep learning face identity features by recovering canonical-view face images

Reconstruction examples from LFW

Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep Learning Identity Preserving Face Space," ICCV 2013.

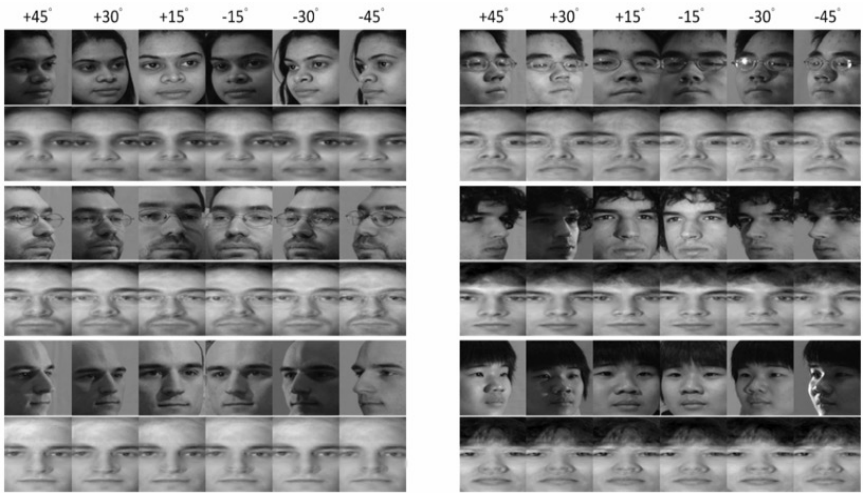
75

- Deep model can disentangle hidden factors through feature extraction over multiple layers
- No 3D model; no prior information on pose and lighting condition
- Model multiple complex transforms
- Reconstructing the whole face is a much stronger supervision than predicting 0/1 class label and helps to avoid over fitting



76

Multi-PIE dataset



77

Comparison on Multi-PIE

	-45°	-30°	-15°	+15°	+30°	+45°	Avg	Pose
LGBP [26]	37.7	62.5	77	83	59.2	36.1	59.3	√
VAAM [17]	74.1	91	95.7	95.7	89.5	74.8	86.9	√
FA-EGFC[3]	84.7	95	99.3	99	92.9	85.2	92.7	x
SA-EGFC[3]	93	**98.7**	99.7	**99.7**	**98.3**	93.6	97.2	√
LE[4] + LDA	86.9	95.5	99.9	**99.7**	95.5	81.8	93.2	x
CRBM[9] + LDA	80.3	90.5	94.9	96.4	88.3	89.8	87.6	x
Wang's	**95.6**	**98.5**	**100.0**	**99.3**	**98.5**	**97.8**	**98.3**	x

[3] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith. Fully automatic pose-invariant face recognition via 3d pose normalization. In *ICCV*, pages 937–944, 2011. 1, 5, 6

[4] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *CVPR*, pages 2707–2714, 2010. 2, 3, 6

[9] G. B. Huang, H. Lee, and E. Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *CVPR*, pages 2518–2525, 2012. 3, 6

[17] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, and S. Shan. Morphable displacement field based image matching for face recognition across pose. In *ECCV*, pages 102–115. 2012. 1, 2, 5, 6

[26] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang. Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition. In *ICCV*, volume 1, pages 786–791, 2005. 5, 6

78

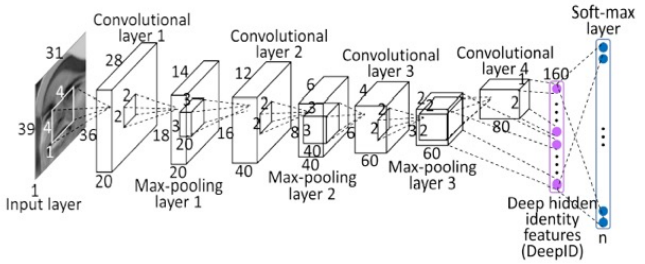
Deep learning 3D model from 2D images, mimicking human brain activities

Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep Learning and Disentangling Face Representation by Multi-View Perception," NIPS 2014.

79

Example 3: deep learning face identity features from predicting 10,000 classes

- At training stage, each input image is classified into 10,000 identities with 160 hidden identity features in the top layers
- The hidden identity features can be well generalized to other tasks (e.g. verification) and identities outside the training set
- As adding the number of classes to be predicted, the generalization power of the learned features also improves



Y. Sun, X. Wang, and X. Tang. Deep Learning Face Representation by Joint Identification-Verification. NIPS, 2014.

80

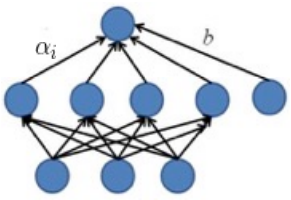
b. Deep Structures vs Shallow Structures (Why deep?)

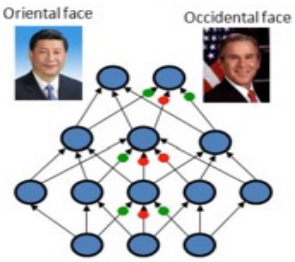
81

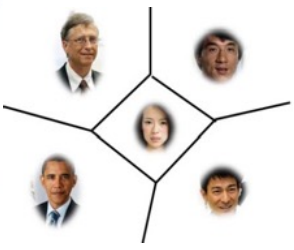
Shallow Structures

- A three-layer neural network (with one hidden layer) can represent any classification function
- Most machine learning tools (such as SVM, boosting, and KNN) can be approximated as neural networks with one or two hidden layers
- Shallow models divide the feature space into regions and match templates in local regions. O(N) parameters are needed to represent N regions

SVM $g(x) = b + \sum_i \alpha_i K(x, x_i)$







82

Deep Machines are More Efficient for Representing Certain Classes of Functions

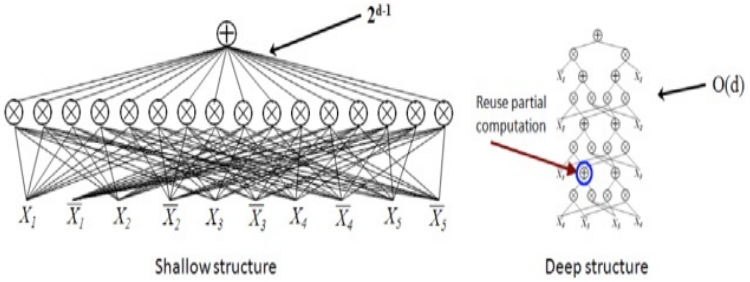
- Theoretical results show that an architecture with insufficient depth can require many more computational elements, potentially exponentially more (with respect to input size), than architectures whose depth is matched to the task
  - (Hastad 1986, Hastad and Goldmann 1991)
- It also means many more parameters to learn…

83

- Take the d-bit parity function as an example

$$(b_1, \ldots, b_d) \in \{0,1\}^d \longmapsto \begin{cases} 1, & \text{if } \sum_{i=1}^{d} b_i \text{ is even} \\ -1, & \text{otherwise} \end{cases}$$

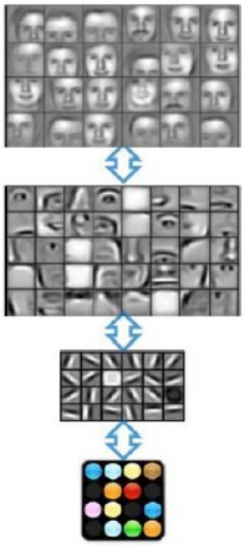
- d-bit logical parity circuits of depth 2 have exponential size (Andrew Yao, 1985)



- There are functions computable with a polynomial-size logic gates circuits of depth k that require exponential size when restricted to depth k -1 (Hastad, 1986)

84

- Architectures with multiple levels provide sharing and re-use of components

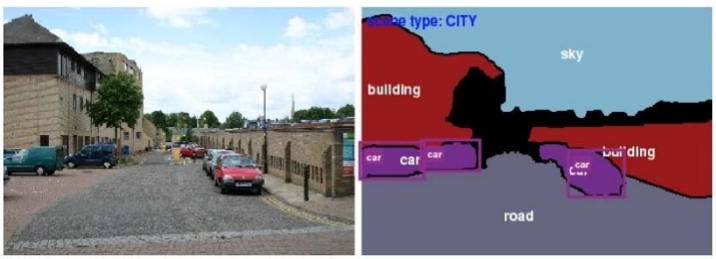


Honglak Lee, NIPS'10

85

Humans Understand the World through Multiple Levels of Abstractions

- We do not interpret a scene image with pixels
  - Objects (sky, cars, roads, buildings, pedestrians) -> parts (wheels, doors, heads) -> texture -> edges -> pixels
  - Attributes: blue sky, red car
- It is natural for humans to decompose a complex problem into sub-problems through multiple levels of representations



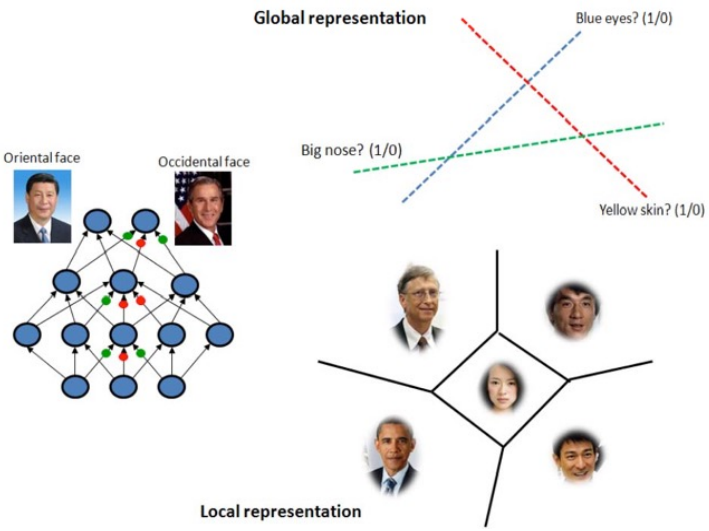
86

Humans Understand the World through Multiple Levels of Abstractions

- Humans learn abstract concepts on top of less abstract ones
- Humans can imagine new pictures by re-configuring these abstractions at multiple levels. Thus our brain has good generalization can recognize things never seen before.
  - Our brain can estimate shape, lighting and pose from a face image and generate new images under various lightings and poses. That's why we have good face recognition capability.

87

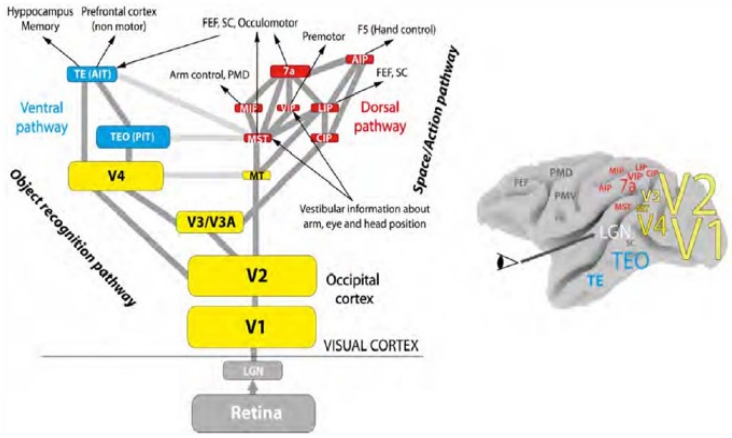
Local and Global Representations



88

Human Brains Process Visual Signals through Multiple Layers

- A visual cortical area consists of six layers (Kruger et al. 2013)

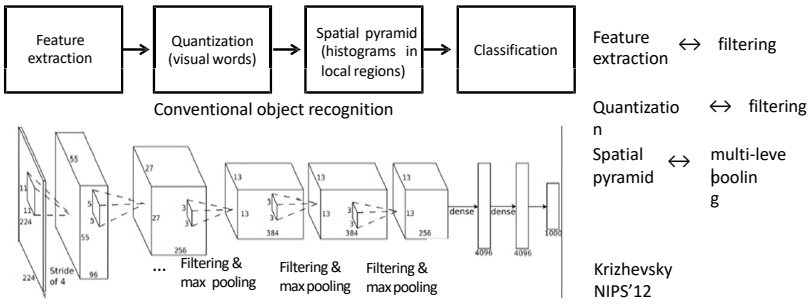


89

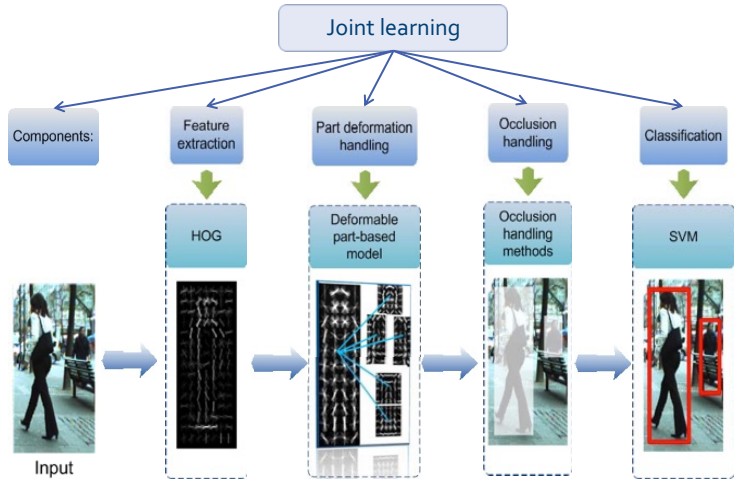
c. Joint Learning vs Separate Learning

90

- Domain knowledge could be helpful for designing new deep models and training strategies
- How to formulate a vision problem with deep learning?
  - Make use of experience and insights obtained in CV research
  - Sequential design/learning vs **joint learning**
  - Effectively train a deep model (layerwise pre-training + fine tuning)



91



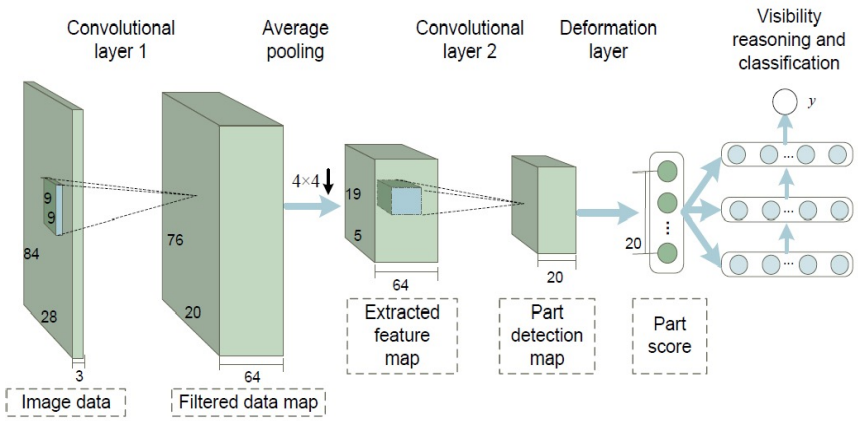
N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. CVPR, 2005. (6000 citations)

P. Felzenszwalb, D. McAlester, and D. Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. CVPR, 2008. (2000 citations)

W. Ouyang, X. Wang. A Discriminative Deep Model for Pedestrian Detect with Occlusion Handling CVPR,2012.

92

Joint Deep Learning Model

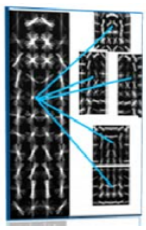


W. Ouyang and X. Wang, "Joint Deep Learning for Pedestrian Detection," Proc. ICCV, 2013.

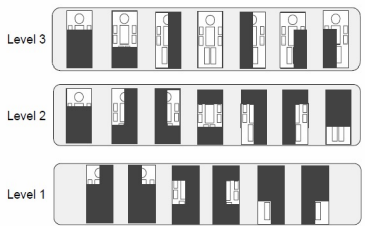
93

Modeling Part Detectors

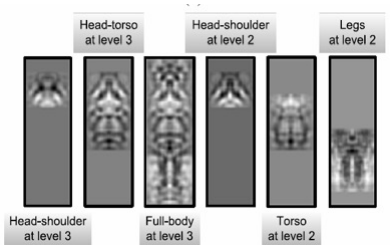
- Design the filters in the second convolutional layer with variable sizes



Part models learned from HOG



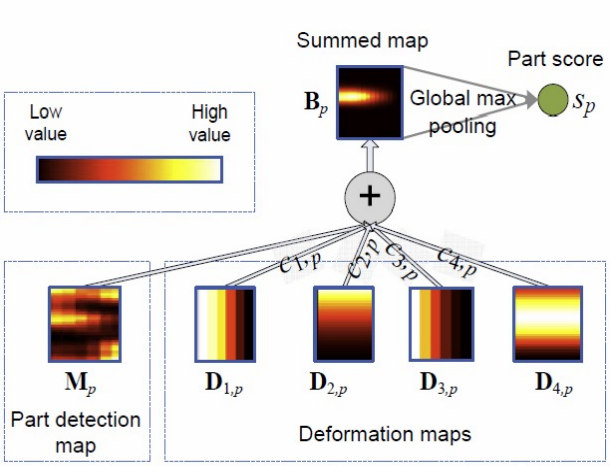
Part models



Learned filtered at the second convolutional layer

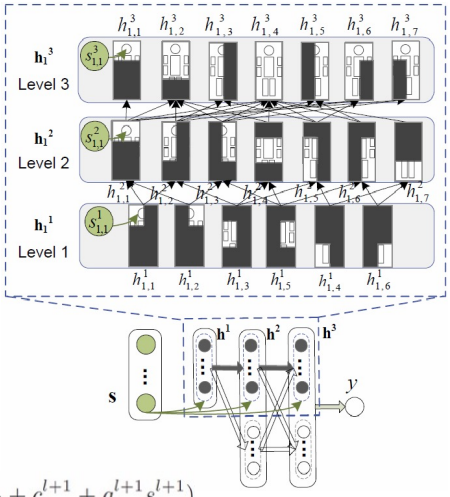
94

Deformation Layer



95

Visibility Reasoning with Deep Belief Net



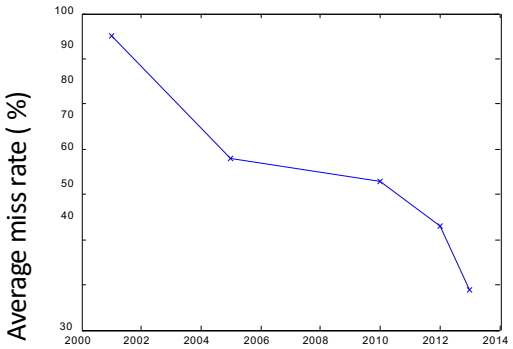
$$\tilde{h}_j^{l+1} = \sigma(\tilde{\mathbf{h}}^{l\mathrm{T}}\mathbf{w}_{*,j}^{l} + c_j^{l+1} + g_j^{l+1}s_j^{l+1})$$

Correlates with part detection score

96

Experimental Results

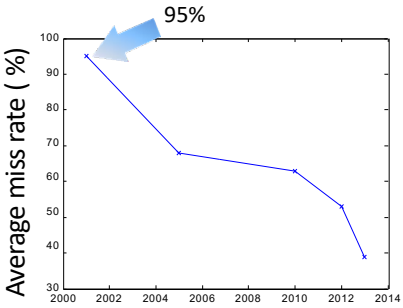
- Caltech – Test dataset (largest, most widely used)



97

Experimental Results

- Caltech – Test dataset (largest, most widely used)

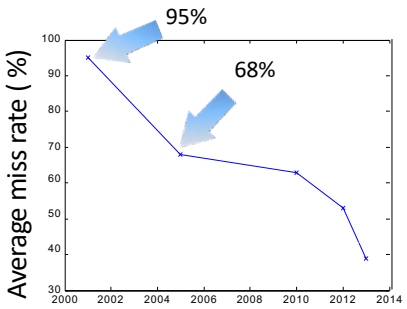


Rapid object detection using a boosted cascade of simple features
P Viola, M Jones - ... Vision and Pattern Recognition, 2001. CVPR ..., 2001 - ieeexplore.ieee.org
Abstract This paper describes a machine learning approach for visual **object detection** which is capable of processing images extremely rapidly and achieving high **detection** rates. This work is distinguished by three key contributions. The first is the introduction of a new ...
Cited by 7647 Related articles All 201 versions Import into BibTeX More

98

Experimental Results

- Caltech – Test dataset (largest, most widely used)

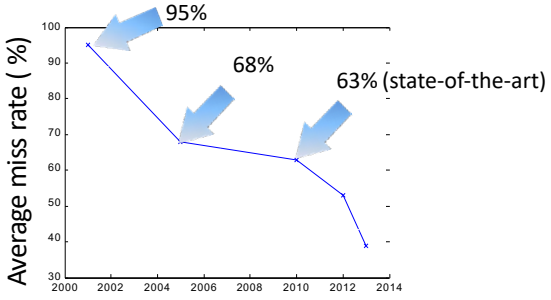


Histograms of oriented **gradients** for **human detection**
N Dalal, B Triggs - ... and Pattern Recognition, 2005. CVPR 2005 ..., 2005 - ieeexplore.ieee.org
... We study the issue of feature sets for **human detection**, showing that lo- cally normalized **Histogram** of Oriented **Gradient** (HOG) de- scriptors provide excellent performance relative to other ex- isting feature sets including wavelets [17,22]. **...**
Cited by 5438 Related articles All 106 versions Import into BibTeX More

99

Experimental Results

- Caltech – Test dataset (largest, most widely used)

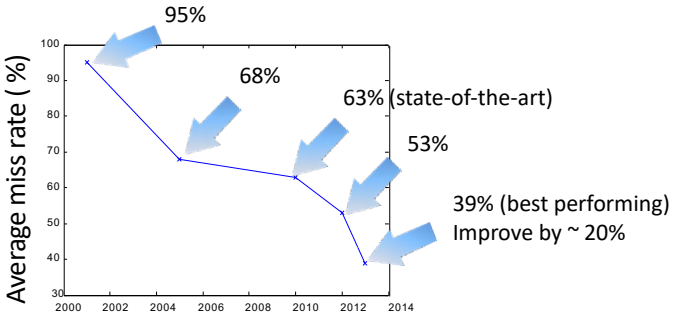


Object detection with **discriminatively trained part-based models**
PF Felzenszwalb, RB Girshick... - Pattern Analysis and ..., 2010 - ieeexplore.ieee.org
Abstract We describe an **object detection** system **based** on mixtures of multiscale deformable **part models**. Our system is able to represent highly variable **object** classes and achieves state-of-the-art results in the PASCAL **object detection** challenges. While **...**
Cited by 964 Related articles All 43 versions Import into BibTeX More

100

Experimental Results

- Caltech – Test dataset (largest, most widely used)



W. Ouyang and X. Wang, "A Discriminative Deep Model for Pedestrian Detection with Occlusion Handling," CVPR 2012.

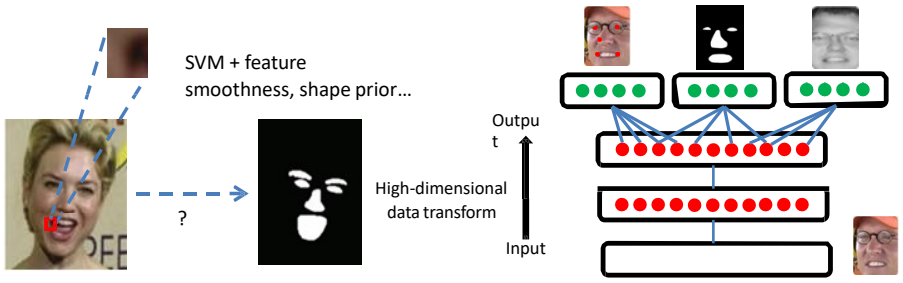
W. Ouyang, X. Zeng and X. Wang, "Modeling Mutual Visibility Relationship in Pedestrian Detection ", CVPR 2013.
W. Ouyang, Xiaogang Wang, "Single-Pedestrian Detection aided by Multi-pedestrian Detection ", CVPR 2013.
X. Zeng, W. Ouyang and X. Wang, „ A Cascaded Deep Learning Architecture for Pedestrian Detection, ICCV 2013.
W. Ouyang and Xiaogang Wang, "Joint Deep Learning for Pedestrian Detection," IEEE ICCV 2013.

101

d. Large learning capacity makes high dimensional data transforms possible

102

- How to make use of the large learning capacity of deep models?
  - **High dimensional data transform**
  - Hierarchical nonlinear representations



103

Face Parsing

P.Luo, X. Wang and X. Tang, "Hierarchical Face Parsing via Deep Learning," CVPR 2012

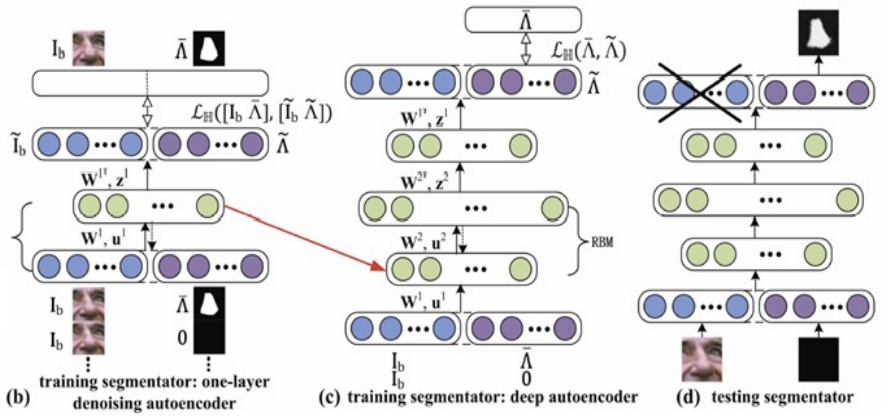
104

Motivations - High dimensional data transform

- Recast face segmentation as a cross-modality data transformation problem
- Cross modality autoencoder
- Data of two different modalities share the same representations in the deep model
- Deep models can be used to learn shape priors for segmentation

105

Training Segmentators



106

107

Summary - High dimensional data transform

- Automatically learning hierarchical feature representations from data and disentangles hidden factors of input data through multi-level nonlinear mappings
- For some tasks, the **expressive power** of deep models **increases exponentially** as their architectures go deep
- Jointly optimize all the components in a vision and crate synergy through close interactions among them
- Benefitting the large learning capacity of deep models, by recasting some classical computer vision challenges - as high-dimensional data transform problems - and solve them from new perspectives

108

Outline

- Introduction to deep learning
- Deep learning for object recognition
- Deep learning for object segmentation
- Deep learning for object detection
- Open questions and future works

111

Outline

- Introduction to deep learning
- Deep learning for object recognition
  - Deep learning for object recognition on ImageNet
  - Deep learning for face recognition
    - Learn identity features from joint verification- identification signals
    - Learn 3D face models from 2D images
- Deep learning for object segmentation
- Deep learning for object detection
- Open questions and future works

112

CNN for Object Recognition on ImageNet

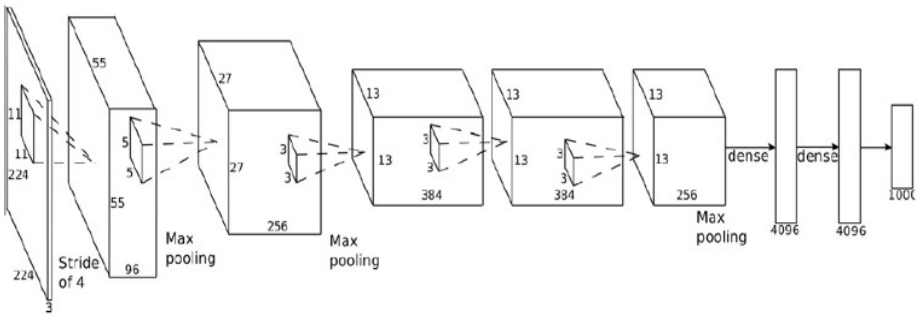
- Krizhevsky, Sutskever, and Hinton, NIPS 2012
- Trained on one million images of **1000 categories** collected from the web with two GPUs; 2GB RAM on each GPU; 5GB of system memory
- Training lasts for one week

Rank	Name	Error rate	Description
1	**U. Toronto**	0.15315	Deep learning
2	U. Tokyo	0.26172	Hand-crafted features and learning models. Bottleneck.
3	U. Oxford	0.26979	
4	Xerox/INRIA	0.27058	

113

Model Architecture

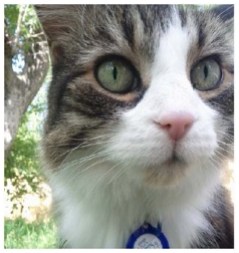
- Max-pooling layers follow 1st, 2nd, and 5th convolutional layers
- The number of neurons in each layer is given by 253440, 186624, 64896, 43264, 4096, 4096, 1000
- 650000 neurons, 60 million parameters, 630 million connections



114

Normalization

- Normalize the input by subtracting the mean image on the training set



Input image (256 x 256)

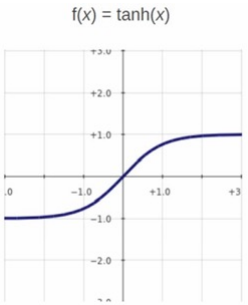
Mean image

Krizhevsky 2012

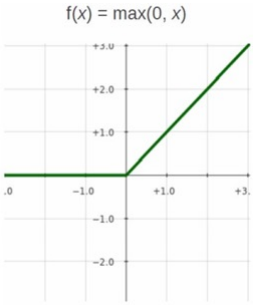
115

Activation Function

- Rectified linear unit leads to sparse responses of neurons, such that weights can be effectively updated with BP



Sigmoid (slow to train)



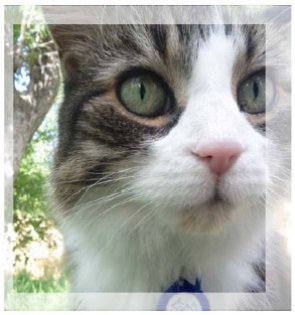
Rectified linear unit (quick to train) √

Krizhevsky 2012

116

Data Augmentation

- The neural net has 60M parameters and it overfits
- Image regions are randomly cropped with shift; their horizontal reflections are also included



Krizhevsky 2012

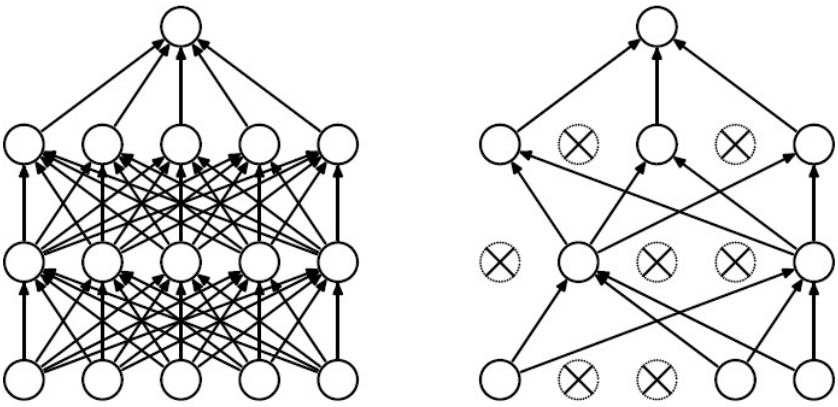
117

Dropout

- Randomly set some input features and the outputs of hidden units as zero during the training process
- Feature co-adaptation: a feature is only helpful when other specific features are present
  - Because of the existence of noise and data corruption, some features or the responses of hidden nodes can be misdetected
- Dropout prevents feature co-adaptation and can significantly **improve the generalization** of the trained network
- Can be considered as another approach to regularization
- It can be viewed as averaging over many neural networks
- Slower convergence

118

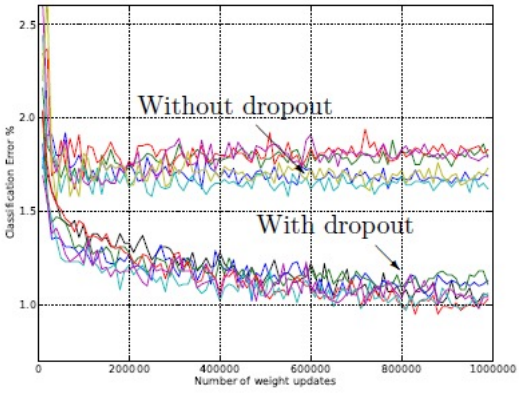
Dropout



(a) Standard Neural Net (b) After applying dropout.

119

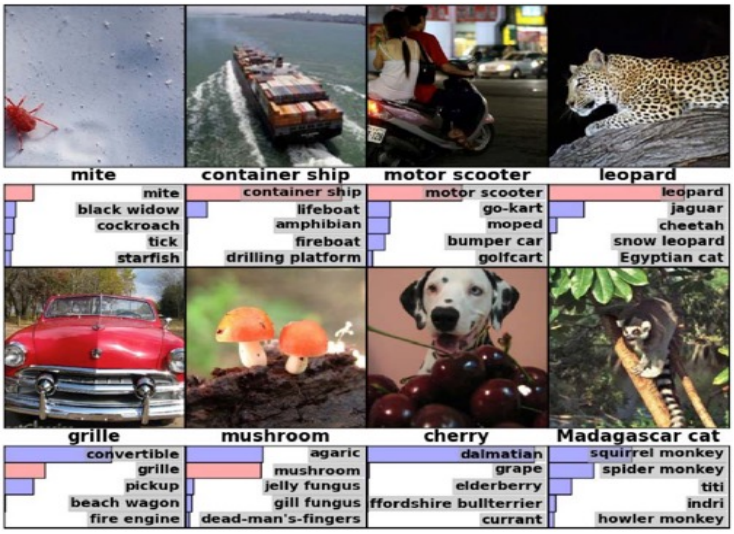
Dropout



Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov; Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research 15 (2014) 1929-1958

120

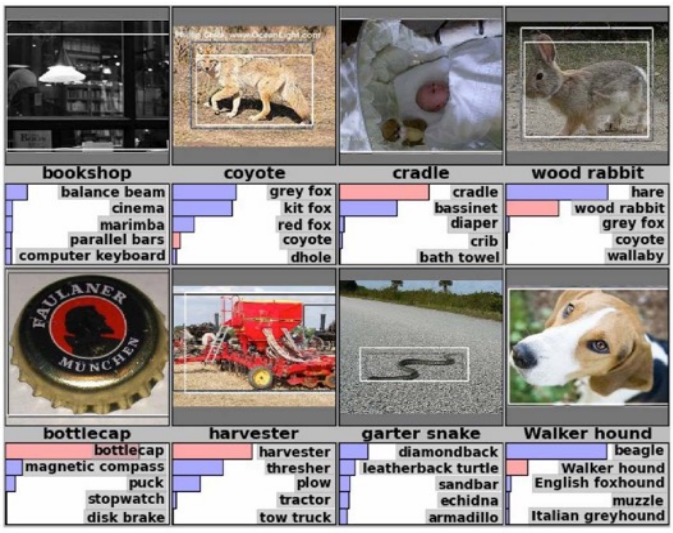
Classification Result



Krizhevsky 2012

121

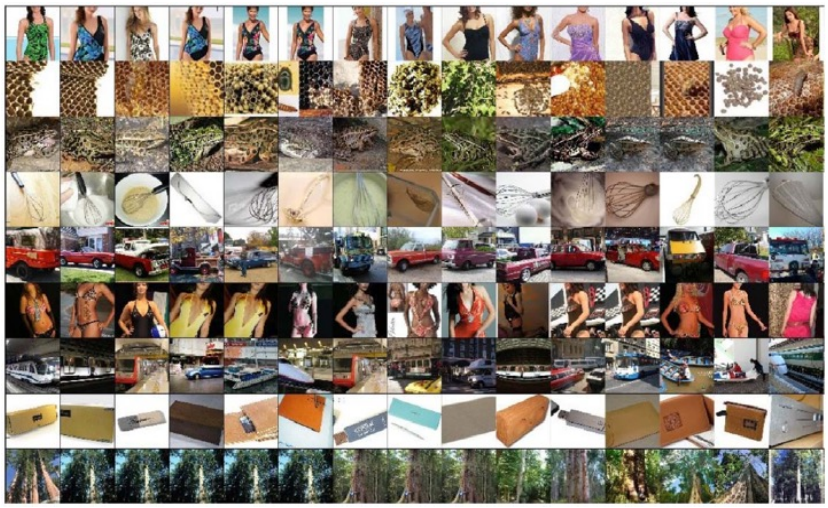
Detection Result



Krizhevsky 2012

122

Image Retrieval

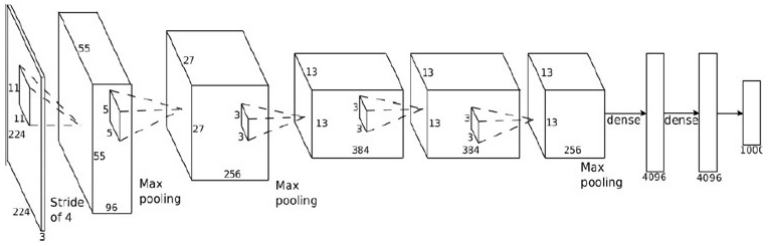


Krizhevsky 2012

123

Adaptation to Smaller Datasets

- Directly use the feature representations learned from ImageNet and replace handcrafted features with them in image classification, scene recognition, fine grained object recognition, attribute recognition, image retrieval (Razavian et al. 2014, Gong et al. 2014)
- Use **ImageNet to pre-train the model (good initialization), and use target dataset to fine-tune it** (Girshick et al. CVPR 2014)
- Fix the bottom layers and **only fine tune the top layers**



124

Outline

- Introduction to deep learning
- Deep learning for object recognition
  - Deep learning for object recognition on ImageNet
  - Deep learning for face recognition
    - Learn identity features from joint verification- identification signals
    - Learn 3D face models from 2D images
- Deep learning for object segmentation
- Deep learning for object detection
- Open questions and future works

125

Deep Learning Results on LFW

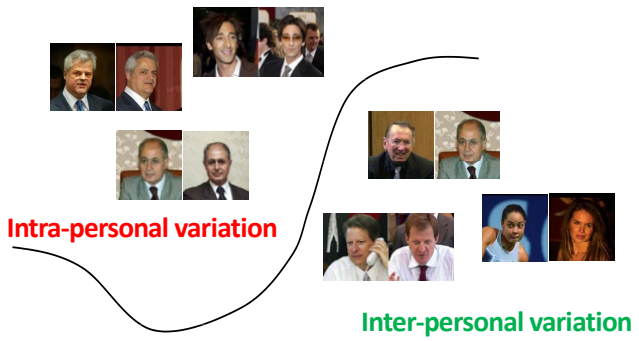
Huang et al. CVPR'12	87%	3	Unsupervised
Sun et al. ICCV'13	92.52%	5	87,628
DeepFace (CVPR'14)	97.35%	6 + 67	7,000,000
Sun et al. (CVPR'14)	97.45%	5	202,599
Sun et al. (arXiv'14)	99.15%	18	202,599

- The first deep learning work on face recognition was done by Huang et al. in 2012. With unsupervised learning, the accuracy was 87%
- Wang's work at ICCV'13 achieved result (92.52%) comparable with state-of-the-art
- Wang's work at CVPR'14 reached **97.45%** close to "human cropped" performance (**97.53%**)
- DeepFace developed by Facebook also at CVPR'14 used 73-point 3D face alignment and 7 million training data (35 times larger than us)
- Wang's work reached **99.15%** close to "human funneled" performance (**99.20%**)

Y. Sun, X. Wang, and X. Tang. Deep Learning Face Representation by Joint Identification-Verification. NIPS, 2014.

126

Eternal Topic on Face Recognition



How to separate the two types of variations?

127

Are they the same person or not?

Nicole Kidman Nicole Kidman

128

Are they the same person or not?

Coo d'Este Melina Kanakaredes

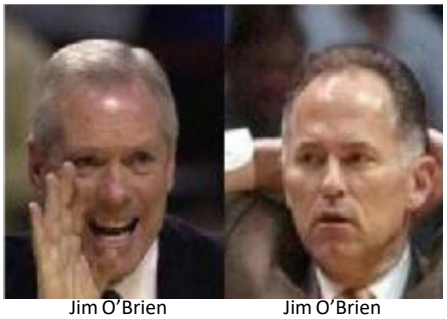
129

Are they the same person or not?

Elijah Wood Stefano Gabbana

130

Are they the same person or not?



Jim O'Brien Jim O'Brien

131

Are they the same person or not?

Jacquline Obradors Julie Taymor

132

- Out of 6000 image pairs on the LFW test set, 51 pairs are misclassified with the deep model
- Randomly mixed them and presented them to 10 subjects for evaluation. Their averaged verification accuracy is 56%, close to random guess (50%)

133

Go Back to the Starting Point

- Eigenface (1992)
- Linear discriminant analysis (LDA) (PAMI'97)
- Bayesian face recognition (PR'00)
- Unified subspace analysis (PAMI'04)

134

Linear Discriminate Analysis

$$W^* = \arg\max_W \frac{\left|W^t S_b W\right|}{\left|W^t S_w W\right|}$$

$$S_b = \sum n_k \left(\bar{x}_k - \bar{x}\right)\left(\bar{x}_k - \bar{x}\right)^t \propto \sum \left(\bar{x}_k - \bar{x}_{k'}\right)\left(\bar{x}_k - \bar{x}_{k'}\right)^t$$

$$S_w = \sum_k \sum_{i \in C_k} \left(x_i - \bar{x}_k\right)\left(x_i - \bar{x}_k\right)^t \propto \sum_{(i,j)\in\Omega} (x_i - x_j)(x_i - x_j)^t$$

P.N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," TPAMI, Vol. 19, pp. 711-720, 1997.

135

$$W^* = \arg\max_W \left|W^t S_b W\right| \quad s.t. \ \left|W^t S_w W\right| = 1$$

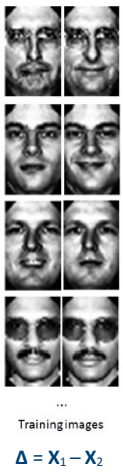
LDA seeks for linear feature mapping which maximizes the distance between class centers under the constraint what the intrapersonal variation is constant

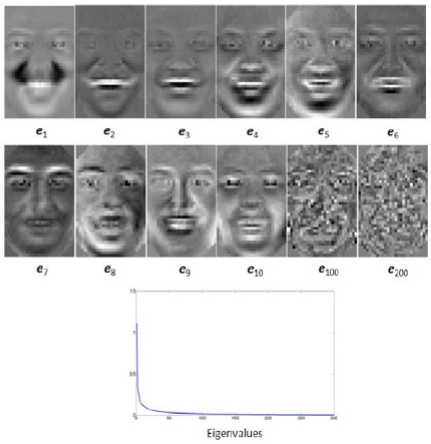
$$f^* = \arg\max_{f^*} \left|W^t S_b W\right| \quad s.t. \ \left|W^t S_w W\right| = 1$$

$$s.t. \sum_{(i,j)\in\Omega} \left|f\left(x_i\right) - f\left(x_j\right)\right|^2 = 1$$

136

Intrapersonal Subspace

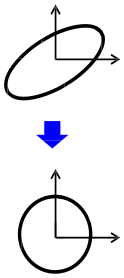




$$\Delta_k = \mathbf{x}_{new} - \bar{\mathbf{x}}_k$$

$$y_{ki} = \mathbf{e}_i^t(\mathbf{x}_{new} - \bar{\mathbf{x}}_k)$$

$$r^2(\Delta_k) = \sum_{i=1}^{d'} y_{ki}^2/\lambda_i$$

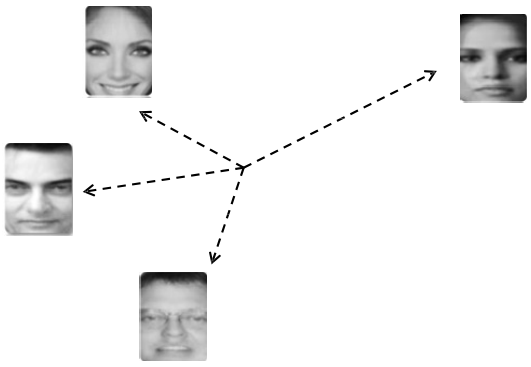


B. Moghaddam, T.Jebara, and A. Pentland, "Bayesian Face Recognition," Pattern Recognition, Vol. 33, pp. 1771-1782, 2000.

137

Scatter Class Centers

- Further do PCA on class centers after reducing intrapersonal variation with whitening



138

Unified Subspace Analysis

- Eigenface: PCA on images to reduce dimensionality and remove noise (when later steps increase intrapersonal difference, some noise could be magnified in wrong directions)
- Bayesianface: PCA on intrapersonal difference vectors to extract the patterns of intrapersonal variations, and depress them by dividing eigenvalues
- Fisherface: PCA on class centers to make them as far as possible and extract identity information

X. Wang and X. Tang, "A Unified Framework for Subspace Face Recognition," TPAMI, Vol. 26, pp. 1222-1228, 2004.

139

Limitations of Existing Approaches

- A lot of information has been lost when calculating the difference $\Delta = X_1 - X_2$

- Linear models with shallow structures cannot separate intra- and inter-personal variations, which are complex, nonlinear, and in high-dimensional image space

140

Deep Learning for Face Recognition

- Extract identity preserving features through hierarchical nonlinear mappings
- Model complex intra- and inter-personal variations with large learning capacity

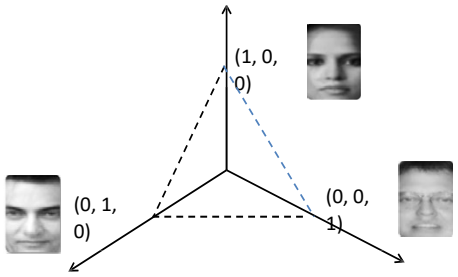
141

Learn Identity Features from Different Supervisory Tasks

- Face identification: classify an image into one of N identity classes
  - multi-class classification problem
- Face verification: verify whether a pair of images belong to the same identity or not
  - binary classification problem

142

Minimize the intra-personal variation under the constraint that the distance between classes is constant (i.e. contracting the volume of the image space without reducing the distance between classes)



$$y = f(x);\ \ g = softmax()$$

$$f^* = \arg\min_f \sum_{(i,j)\in\Omega} \left\| f(x_i) - f(x_j) \right\|^2$$

$$\left| g(f(x_i)) - g(f(x_j)) \right| = 1,\ \ label(x_i) \neq label(x_j)$$

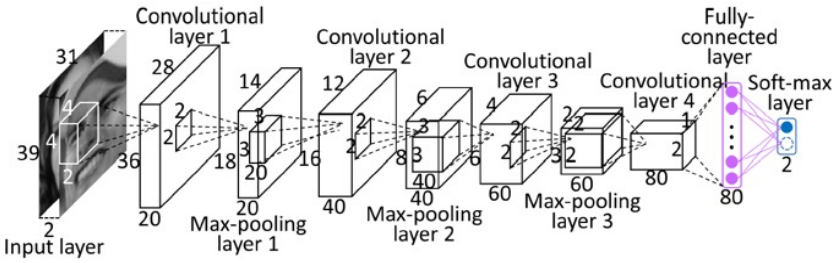
143

Learn Identity Features with Verification Signal

- Extract relational features with learned filter pairs

$$y^j = f\left(b^j + k^{1j} * x^1 + k^{2j} * x^2\right)$$

- These relational features are further processed through multiple layers to extract global features
- The fully connected layer can be used as features to combine with multiple ConvNets

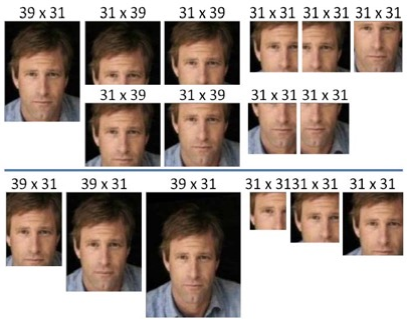


Y. Sun, X. Wang, and X. Tang, "Hybrid Deep Learning for Computing Face Similarities," Proc. ICCV, 2013.

144

Generate Multiple CNNs

- 10 face regions, 3 scales, color/gray and 8 modes
- Base on three-point alignment

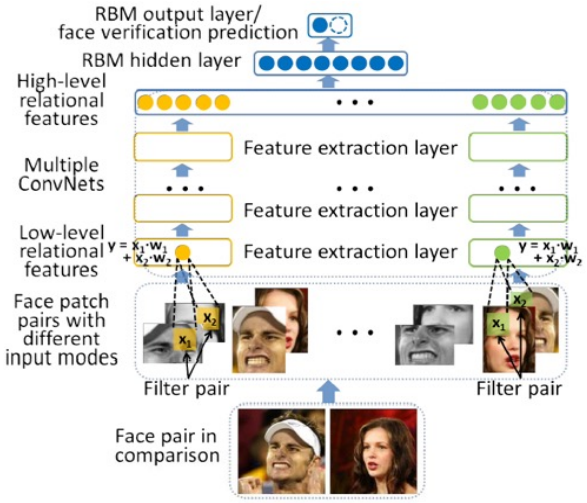


Regions and scales

modes

145

RBM Combines Features Extracted by Multiple ConvNets



146

Results on LFW

- Outside training data: the CelebFaces dataset has 87,628 face images of 5,436 celebrities. Its identities have no overlap with LFW

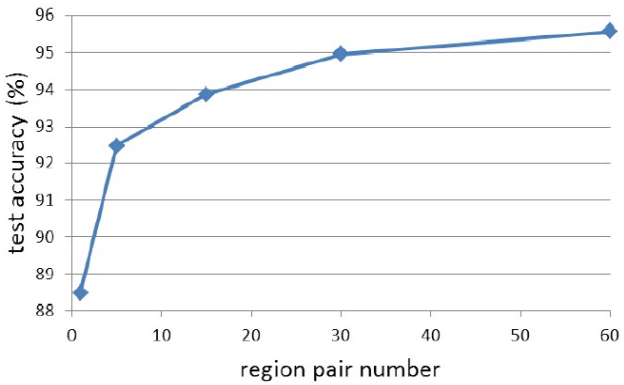
	hid	hid+out	out
dimension	38,400	38,880	480
each dim (%)	60.25	60.58	86.63
PCA+LDA (%)	94.55	94.42	93.41
SVM linear (%)	95.12	95.04	93.45
SVM rbf (%)	94.95	94.89	94.00
classRBM (%)	**95.56**	95.32	93.79

Taking the last hidden layer (hid) as features for combination is more effective than using the output of CNNs (out)

147

Results on LFW

- More regions improve performance



148

Results on LFW

- Fine tuning RBM and ConvNets improves the performance
- Averaging 5 RBMs (each is trained with a randomly generated training set) can improves performance

	LFW (%)	CelebFaces (%)
Single ConvNet	85.05	88.46
RBM	93.45	95.56
Fine-tuning	93.58	96.60
Model averaging	**93.83**	**97.08**

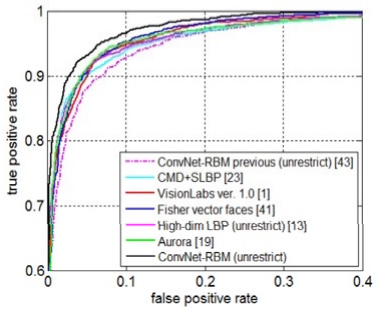
LFW: only using training images from LFW with unrestricted protocol CelebFaces: using CelebFaces as training set without training images from LFW

149

Results on LFW

- Unrestricted protocol without outside training data

Method	Accuracy (%)
ConvNet-RBM previous [43]	91.75 ± 0.48
VMRS [3]	92.05 ± 0.45
CMD+SLBP [23]	92.58 ± 1.36
VisionLabs ver. 1.0 [1]	92.90 ± 0.31
Fisher vector faces [41]	93.03 ± 1.05
High-dim LBP [13]	93.18 ± 1.07
Aurora [19]	93.24 ± 0.44
ConvNet-RBM	$\mathbf{93.83 \pm 0.52}$

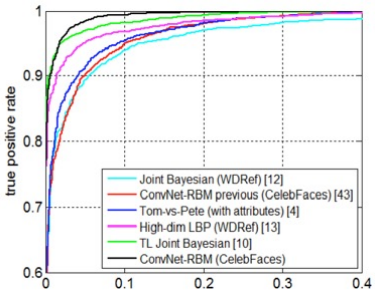


150

Results on LFW

- Unrestricted protocol using outside training data

Method	Accuracy (%)
Joint Bayesian [12]	92.42 ± 1.08
ConvNet-RBM previous [43]	92.52 ± 0.38
Tom-vs-Pete (with attributes) [4]	93.30 ± 1.28
High-dim LBP [13]	95.17 ± 1.13
TL Joint Bayesian [10]	96.33 ± 1.08
ConvNet-RBM	$\mathbf{97.08 \pm 0.28}$



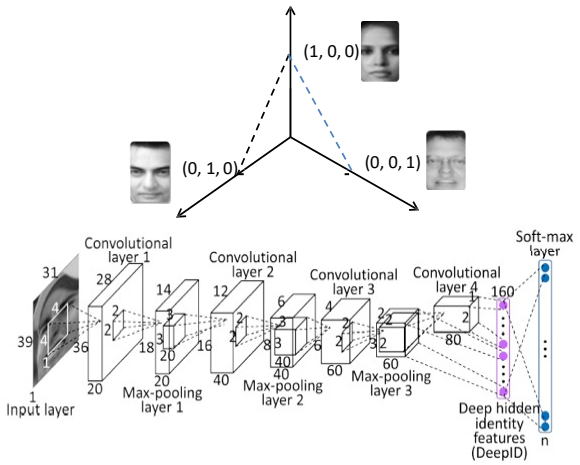
151

Summary of the Results

- Use the last hidden layer instead of the output of CNNs as features
- Fusion of features from more face regions (CNNs) improves the performance
- Fine tuning RBM and CNNs improves performance
- Averaging the outputs of multiple RBMs improves the performance
- Drawbacks: computational cost is high and features cannot be computed offline

152

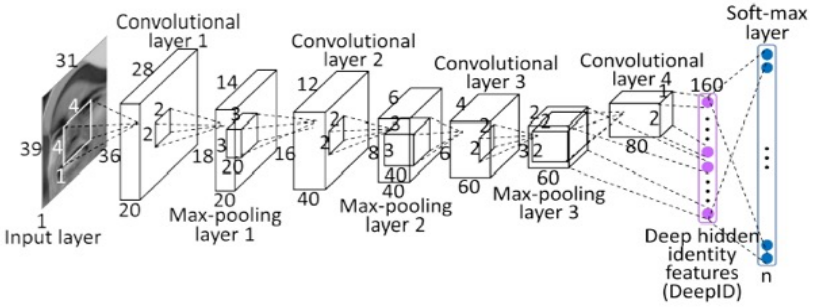
Learn Identity Features with Identification Signal



Y. Sun, X. Wang, and X. Tang, "Deep Learning Face Representation from Predicting 10,000 classes," Proc. CVPR, 2014.

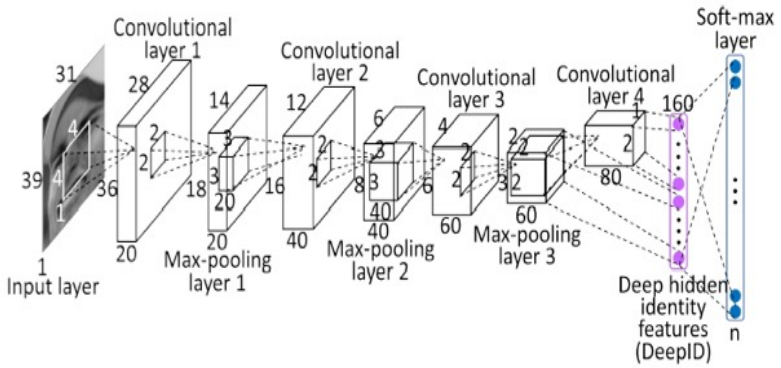
153

- During training, each image is classified into 10,000 identities with 160 identity features in the top layer
- These features keep rich inter-personal variations
- Features from the last two convolutional layers are effective
- The hidden identity features can be well generalized to other tasks (e.g. verification) and identities outside the training set



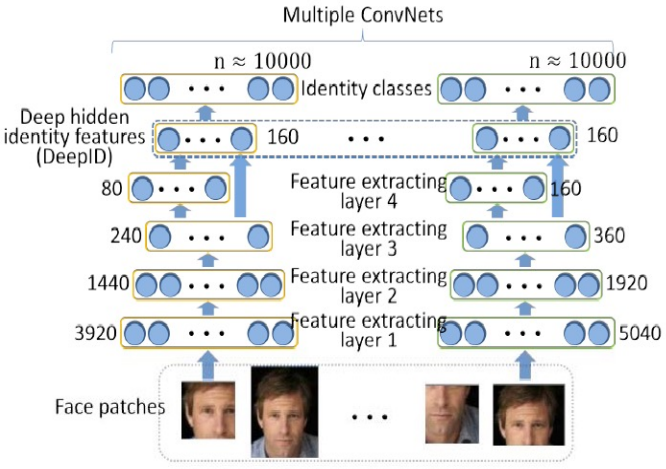
154

- High-dimensional prediction is more challenging, but also adds stronger supervision to the network
- As adding the number of classes to be predicted, the generalization power of the learned features also improves



155

Extract Features from Multiple ConvNets



156

Learn Identity Features with Identification Signal

- After combining hidden identity features from multiple CovNets and further reducing dimensionality with PCA, each face image has 150-dimenional features as signature
- These features can be further processed by other classifiers in face verification. Interestingly, we find Joint Bayesian is more effective than cascading another neural network to classify these features

157

Result on LFW

- Enlarge CelebFaces dataset to CelebFaces+, which include 202,599 images of 10,117 celebrities. CelebFaces+ has no overlap with LFW on identities

Method	Accuracy (%)	No. of points	No. of images	Feature dimension
Joint Bayesian [8]	92.42 (o)	5	99,773	2000 × 4
ConvNet-RBM [31]	92.52 (o)	3	87,628	N/A
CMD+SLBP [17]	92.58 (u)	3	N/A	2302
Fisher vector faces [29]	93.03 (u)	9	N/A	128 × 2
Tom-vs-Pete classifiers [2]	93.30 (o+r)	95	20,639	5000
High-dim LBP [9]	95.17 (o)	27	99,773	2000
TL Joint Bayesian [6]	96.33 (o+u)	27	99,773	2000
DeepFace [32]	97.25 (o+u)	6 + 67	4,400,000 + 3,000,000	4096 × 4
DeepID on CelebFaces	**96.05** (o)	5	87,628	150
DeepID on CelebFaces+	**97.05** (o)	5	202,599	150
DeepID on CelebFaces+ with transfer	**97.45** (o+u)	5	202,599	150

"o" denotes using outside training data, however, without using training data from LFW

"o+u" denotes using outside training data and LFW data in the unrestricted protocol for training

158

Joint Identification-Verification Signals

- Every two feature vectors extracted from the same identity should are close to each other

$$\text{Verif}(f_i, f_j, y_{ij}, \theta_{ve}) = \begin{cases} \frac{1}{2}\|f_i - f_j\|_2^2 & \text{if } y_{ij} = 1 \\ \frac{1}{2}\max\left(0, m - \|f_i - f_j\|_2\right)^2 & \text{if } y_{ij} = -1 \end{cases}$$

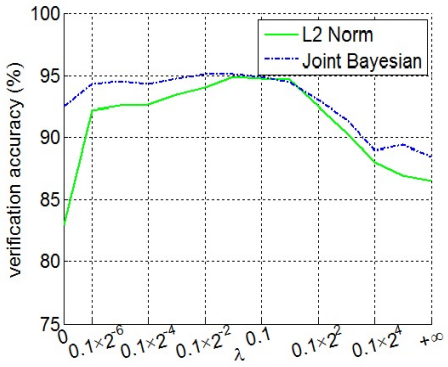
f_i and f_j are feature vectors extracted from two face images in comparison

y_{ij} = 1 means they are from the same identity; y_{ij} = -1 means different identities

m is a margin to be learned

159

Balancing Identification and Verification Signals with Parameter λ

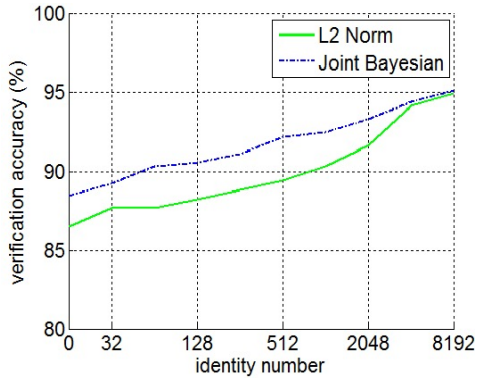


λ = 0: only identification signal
λ = +∞: only verification signal

160

Rich Identity Information Improves Feature Learning

- Face verification accuracies with the number of training identities



161

Final Result

- 25 face regions at different scales and locations around landmarks are selected to build 25 neural networks
- All the 160 X 25 hidden identity features are further compressed into a 180-dimensional feature vector with PCA as a signature for each image
- With a single Titan GPU, the feature extraction process takes 35ms per image

162

Final Result

Methods	High-dim LBP [1]	TL Joint Bayesian [2]	DeepFace [3]	DeepID [4]	DeepIV
Accuracy (%)	95.17	96.33	97.35	97.45	99.15

[1] Chen, Cao, Wen, and Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. CVPR, 2013.

[2] Cao, Wipf, Wen, Duan, and Sun. A practical transfer learning algorithm for face verification. ICCV, 2013.

[3] Taigman, Yang, Ranzato, and Wolf. DeepFace: Closing the gap to human-level performance in face verification. *CVPR,* 2014.

[4] Sun, Wang, and Tang. Deep learning face representation from predicting 10,000 classes. *CVPR,* 2014.

163

Unified subspace analysis

- Identification signal is in S_b; verification signal is in S_w
- Maximize distance between classes under constraint that intrapersonal variation is constant
- Linear feature mapping

Joint deep learning

- Learn features by joint identification-verification
- Minimize intra-personal variation under constraint that the distance between classes is constant
- Hierarchical nonlinear feature extraction
- Generalization power increases with more training identities

- Need to be careful when magnifying the inter-personal difference; Unsupervised learning many be a good choice to remove noise

We still do not know the limit of deep learning yet

164

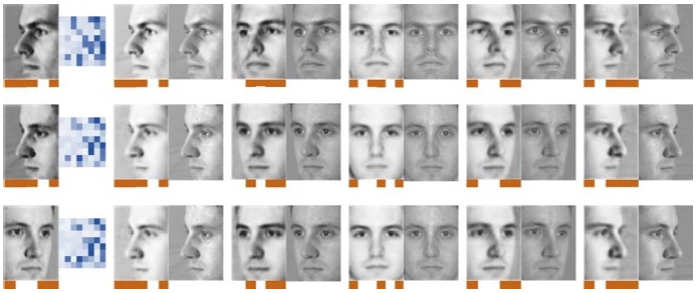
Outline

- Introduction to deep learning
- Deep learning for object recognition
  - Deep learning for object recognition on ImageNet
  - Deep learning for face recognition
    - Learn identity features from joint verification- identification signals
    - Learn 3D face models from 2D images
- Deep learning for object segmentation
- Deep learning for object detection
- Open questions and future works

165

Deep Learning Multi-view Representation from 2D Images

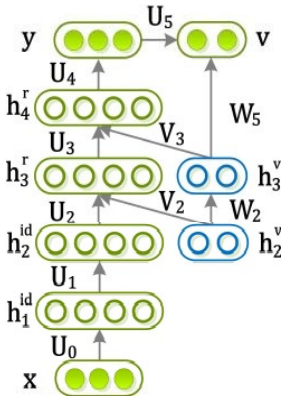
- Inspired by brain behaviors
- Identity and view represented by different sets of neurons
- Given an image under arbitrary view, its viewpoint can be estimated and its full spectrum of views can be reconstructed



Z. Zhu, P.Luo, X. Wang, and X. Tang, "Deep Learning and Disentangling Face Representation by Multi-View Perception," NIPS 2014.

166

Deep Learning Multi-view Representation from 2D Images



x and y are input and ouput images of the same identity but in different views;

v is the view label of the output image;

h^{id} are neurons encoding identity features

h^v are neurons encoding view features

h^r are neurons encoding features to reconstruct the output images

167

	Avg.	0°	−15°	+15°	−30°	+30°	−45°	+45°	−60°	+60°
Raw Pixels+LDA	36.7	81.3	59.2	58.3	35.5	37.3	21.0	19.7	12.8	7.63
LBP [1]+LDA	50.2	89.1	77.4	79.1	56.8	55.9	35.2	29.7	16.2	14.6
Landmark LBP [6]+LDA	63.2	94.9	83.9	82.9	71.4	68.2	52.8	48.3	35.5	32.1
CNN+LDA	58.1	64.6	66.2	62.8	60.7	63.6	56.4	57.9	46.4	44.2
FIP [28]+LDA	72.9	94.3	91.4	90.0	78.9	82.5	66.1	62.0	49.3	42.5
RL [28]+LDA	70.8	94.3	90.5	89.8	77.5	80.0	63.6	59.5	44.6	38.9
MTL+RL+LDA	**74.8**	**93.8**	**91.7**	**89.6**	**80.1**	**83.3**	**70.4**	**63.8**	51.5	50.2
$MVP_{h_1^{id}}$+LDA	61.5	92.5	85.4	84.9	64.3	67.0	51.6	45.4	35.1	28.3
$MVP_{h_2^{id}}$+LDA	**79.3**	**95.7**	**93.3**	**92.2**	**83.4**	**83.9**	**75.2**	**70.6**	**60.2**	**60.0**
$MVP_{h_3^v}$+LDA	72.6	91.0	86.7	84.1	74.6	74.2	68.5	**63.8**	**55.7**	**56.0**
$MVP_{h_4^r}$+LDA	62.3	83.4	77.3	73.1	62.0	63.9	57.3	53.2	44.4	46.9

Face recognition accuracies across views and illuminations on the Multi-PIE dataset. The first and the second best performances are in bold.

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *TPAMI*, 28:2037–2041, 2006.

[6] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *CVPR*, 2013.

[28] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity preserving face space. In *ICCV*, 2013.

168

Deep Learning Multi-view Representation from 2D Images

- Interpolate and predict images under viewpoints unobserved in the training set

The training set only has viewpoints of 0°, 30°, and 60°. (a): the reconstructed images under 15° and 45° when the input is taken under 0°. (b) The input images are under 15° and 45°.

169

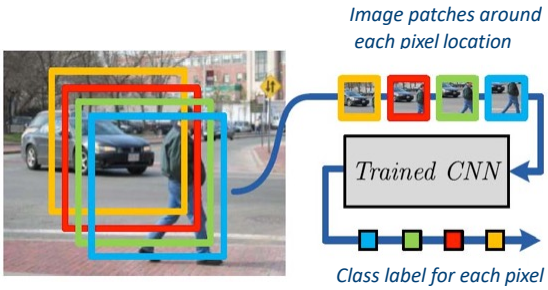
Outline

- Introduction to deep learning
- Deep learning for object recognition
- Deep learning for object segmentation
- Deep learning for object detection
- Open questions and future works

170

Pixelwise Classification

- Image patches centered at each pixel are used as the input of a CNN, and the CNN predicts a class label for each pixel

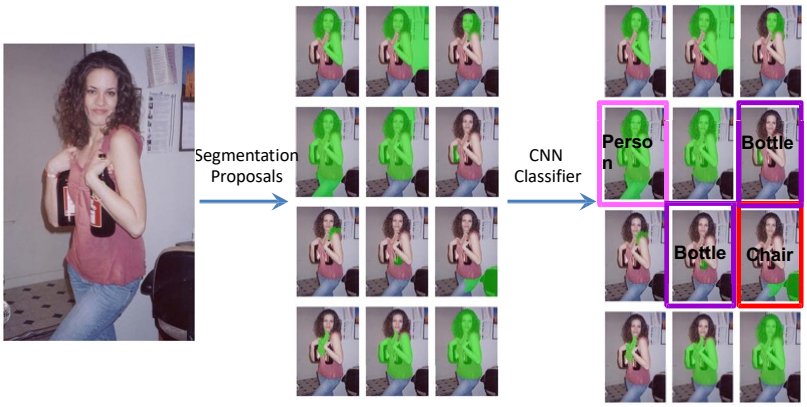


Farabet et al. TPAMI 2013 Pinheiro and Collobert ICML 2014

171

Classify Segmentation Proposal

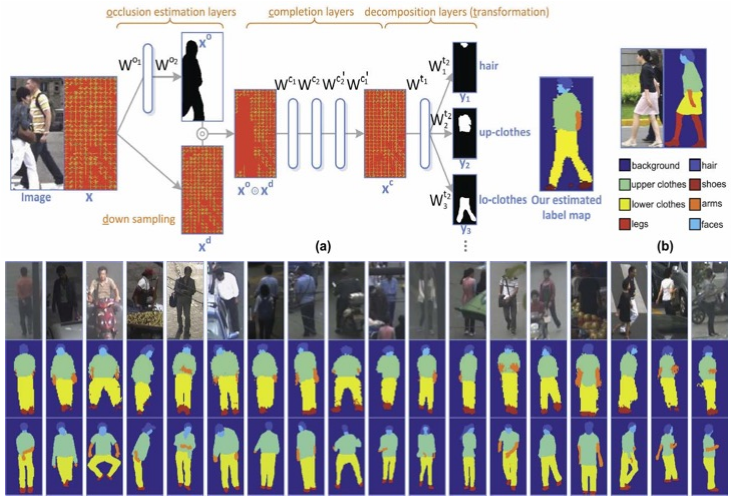
- Determines which segmentation proposal can best represent objects on interest



R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation" CVPR 2014

172

Direct Predict Segmentation Maps



P. Luo, X. Wang, and X. Tang, "Pedestrian Parsing via Deep Decompositional Network," ICCV 2013.

173

Summary

- Deep learning significantly outperforms conventional vision systems on large scale image classification
- Feature representation learned from ImageNet can be well generalized to other tasks and datasets
- In face recognition, identity preserving features can be effectively learned by joint identification-verification signals
- 3D face models can be learned from 2D images; identity and pose information is encoded by different sets of neurons
- We still do not see the limit of the deep model yet, as the size of the training set increases
- In segmentation, larger patches lead to better performance because of the large learning capacity of deep models. It is also possible to directly predict the segmentation map.

174

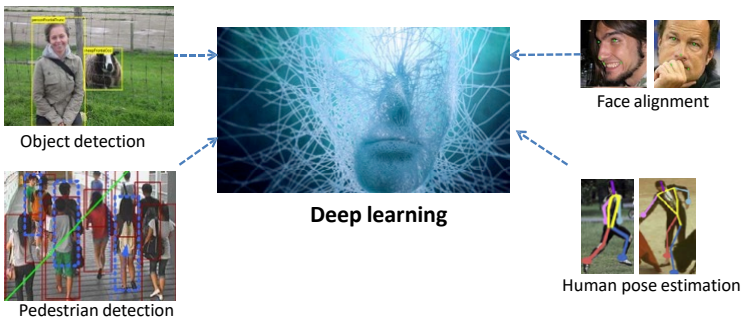
Outline

- Introduction to deep learning
- Deep learning for object recognition
- Deep learning for object segmentation
- Deep learning for object detection
- Open questions and future works

177

Deep Learning for Object Detection

- Pedestrian Detection
- Human part localization
- General object detection



178

2"How to"s

- How to effectively train a deep model
  - Data augmentation
  - Label more data
  - Pre-train on large-scale related data (RCNN)
  - Layerwise pre-training + fine tuning (Multi-stage)
- How to formulate a vision problem with deep learning
  - Tune hyper-parameters, e.g. number of hidden nodes, filter size, number of layers, activation function, dropout …
  - Make use of experience and insights obtained in CV research
    - Sequential design/learning vs joint learning
    - Contextual information (Multi-stage, face, human pose)
    - Background clutter removal (SDN)

179

2"How to"s

- How to effectively train a deep model
  - Data augmentation
  - Label more data
  - Pre-train on large-scale related data (RCNN)
  - Layerwise pre-training + fine tuning (Multi-stage)
- How to formulate a vision problem with deep learning
  - Tune hyper-parameters, e.g. number of hidden nodes, filter size, number of layers, activation function, dropout …
  - Make use of experience and insights obtained in CV research
    - Sequential design/learning vs joint learning
    - Contextual information (Multi-stage, face, human pose)
    - Background clutter removal (SDN)

180

2"How to"s

- How to effectively train a deep model
  - Data augmentation
  - Label more data
  - **Pre-train on large-scale related data (RCNN*)**
  - **Layerwise pre-training + fine tuning (Multi-stage)**

How to formulate a vision problem with deep learning

- Tune hyper-parameters, e.g. number of hidden nodes, number of layers, activation function, dropout.
- **Make use of experience and insights obtained in CV research**
  - Sequential design/learning vs **joint learning**
  - Contextual information (Multi-stage, face, human pose)
  - Background clutter removal (SDN)
  - Short and long range temporal relationship (Action recognition)

*RCNN = Region-based Convolutional Neural Networks

181

2"How to"s

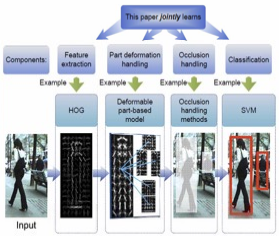
- How to effectively train a deep model
  - Data augmentation
  - Label more data
  - Pre-train on large-scale related data (RCNN)
  - Layerwise pre-training + fine tuning (Multi-stage)

How to formulate a vision problem with deep learning

- Tune hyper-parameters, e.g. number of hidden nodes, number of layers, activation function, dropout.
- Make use of experience and insights obtained in CV research
  - **Sequential design/learning vs joint learning**
  - Contextual information (Multi-stage, face, human pose)
  - Background clutter removal (SDN)
  - Short and long range temporal relationship (Li fei-fei and Yu kai's works)

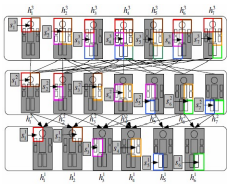
182

Pedestrian Detection

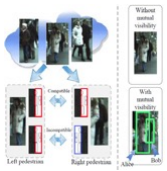


Ouyang et al. ICCV'13

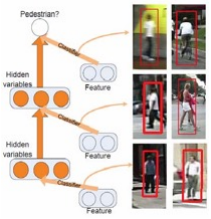
Improve state-of-the-art average miss detection rate on the largest Caltech dataset from 63% to 38%



Ouyang et al. CVPR'12



Ouyang et al. CVPR'13



Zeng et al. ICCV'13

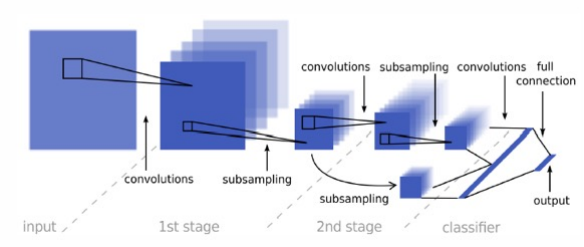
Luo et al. CVPR'14

183

Joint Learning vs Separate Learning

184

What if we treat an existing deep model as a black box in pedestrian detection?

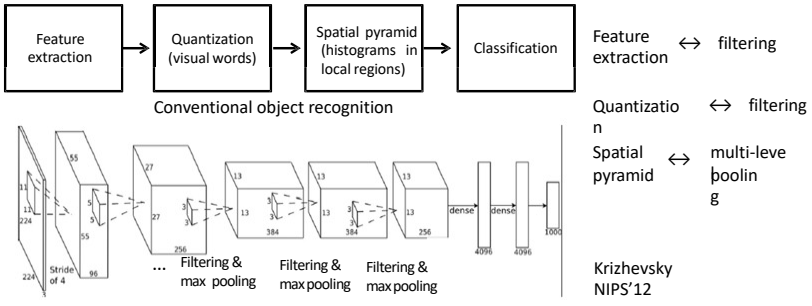


ConvNet−U−MS

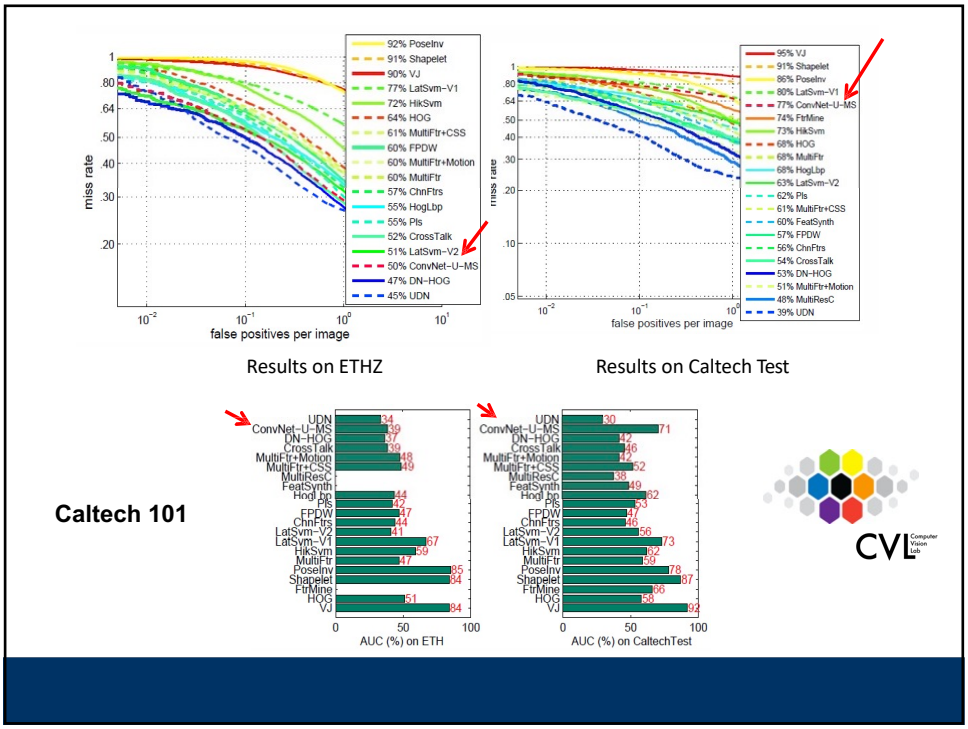
– Sermnet, K. Kavukcuoglu, S. Chintala, and LeCun, "Pedestrian Detection with Unsupervised Multi-Stage Feature Learning," CVPR 2013.

185

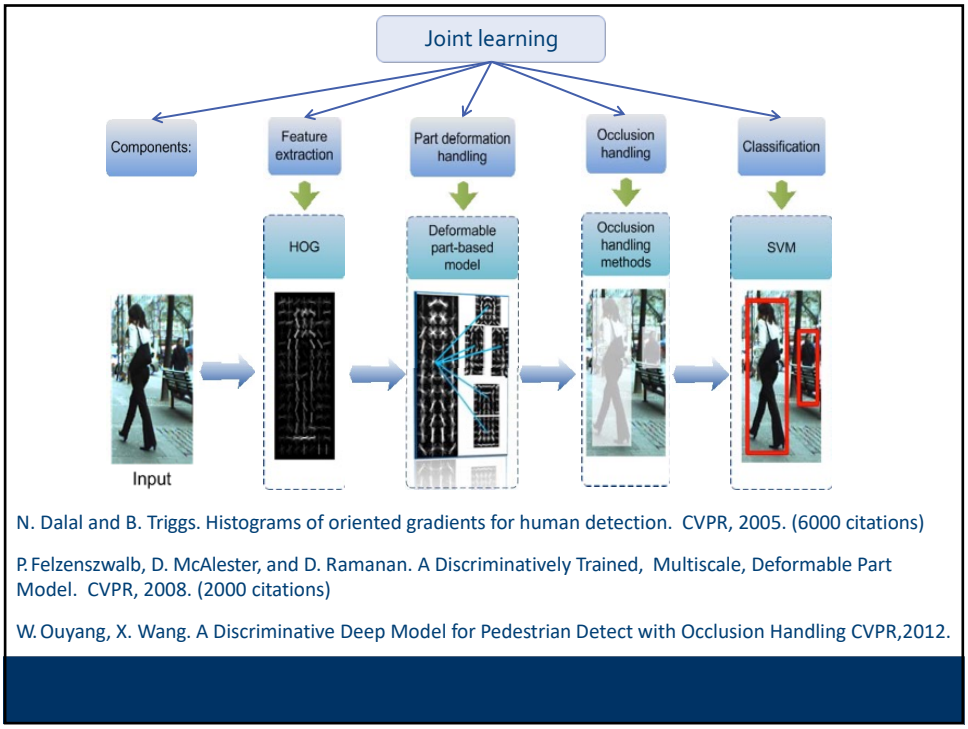
- Domain knowledge could be helpful for designing new deep models and training strategies
- How to formulate a vision problem with deep learning?
  - Make use of experience and insights obtained in CV research
  - Sequential design/learning vs **joint learning**
  - Effectively train a deep model (layerwise pre-training + fine tuning)



186



187



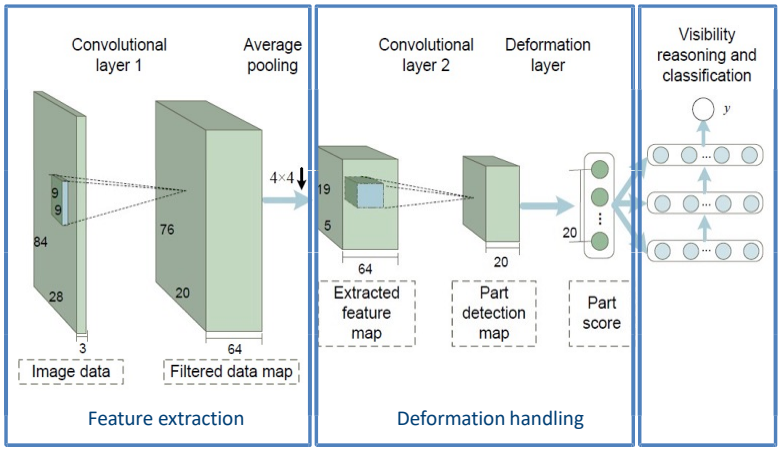
N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. CVPR, 2005. (6000 citations)

P. Felzenszwalb, D. McAlester, and D. Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. CVPR, 2008. (2000 citations)

W. Ouyang, X. Wang. A Discriminative Deep Model for Pedestrian Detect with Occlusion Handling CVPR,2012.

188

A Joint Deep Learning Model

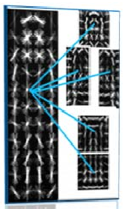


W. Ouyang and X. Wang, "Joint Deep Learning for Pedestrian Detection," Proc. ICCV, 2013.

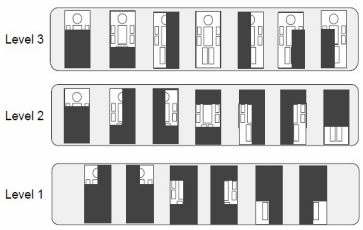
189

Modeling Part Detectors

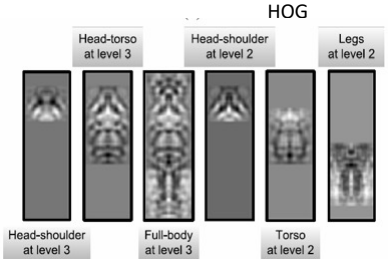
- Design the filters in the second convolutional layer with variable sizes



Part models learned from HOG



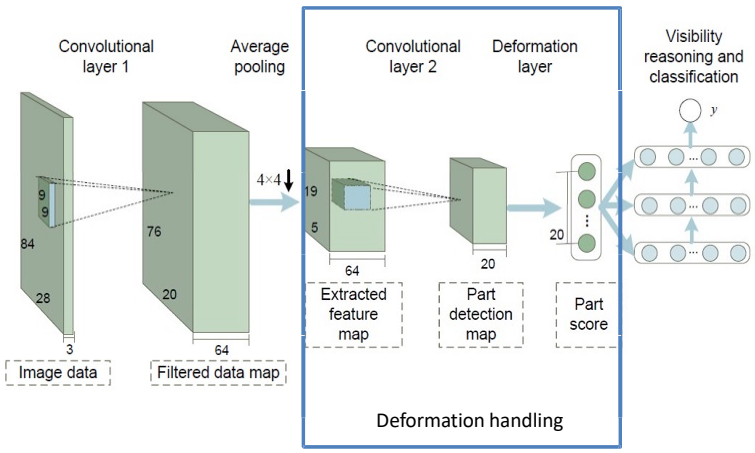
Part models



Learned filtered at the second convolutional layer

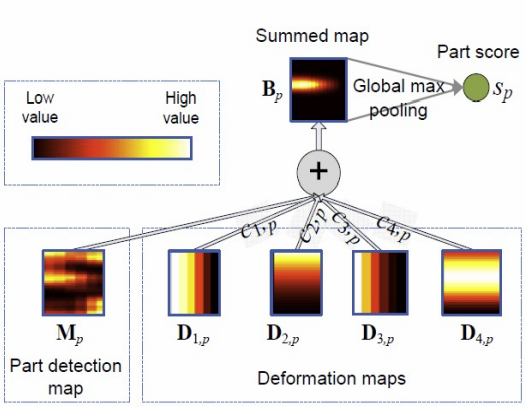
190

A Joint Deep Learning Model



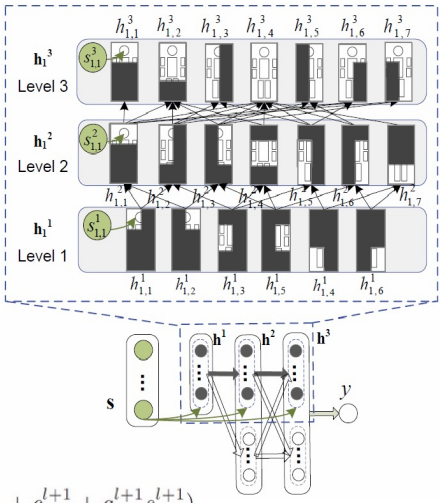
191

Deformation Layer



192

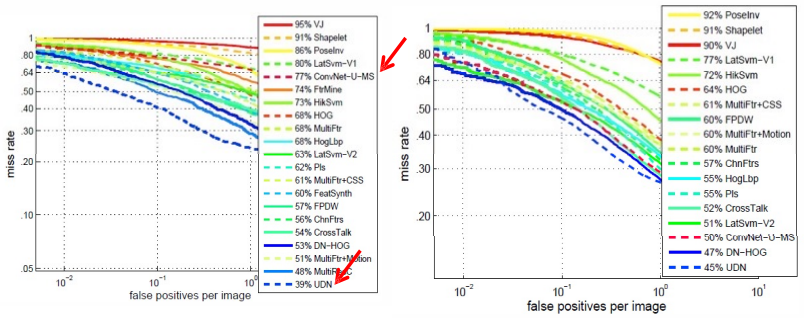
Visibility Reasoning with Deep Belief Net



$$\tilde{h}_j^{l+1} = \sigma(\tilde{\mathbf{h}}^{l\mathrm{T}}\mathbf{w}_{*,j}^{l} + c_j^{l+1} + g_j^{l+1}s_j^{l+1})$$

Correlates with part detection score

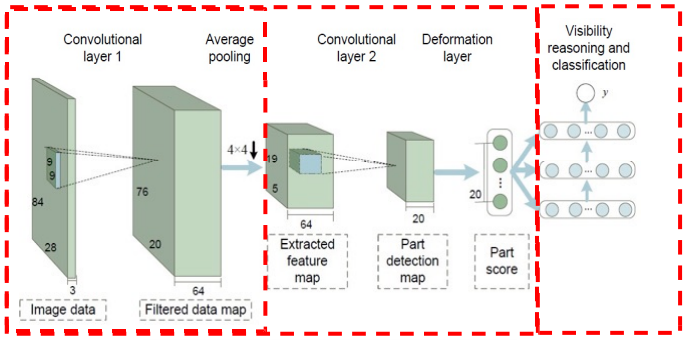
193

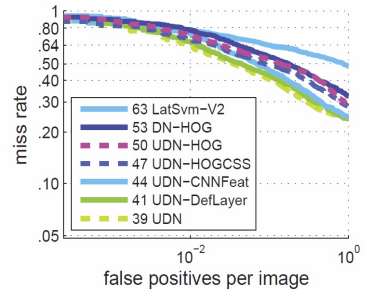


Results on Caltech Test

Results on ETHZ

194





DN-HOG
UDN-HOG
UDN-HOGCSS
UDN-CNNFeat
UDN-DefLayer

195

2 "How to"s

- How to effectively train a deep model
  - Data augmentation
  - Label more data
  - Pre-train on large-scale related data (RCNN)
  - **Layerwise pre-training + fine tuning (Multi-stage)**
- How to formulate a vision problem with deep learning
  - Tune hyper-parameters, e.g. number of hidden nodes, number of layers, activation function, dropout.
  - Make use of experience and insights obtained in CV research
    - **Sequential design/learning vs joint learning**
    - **Contextual information** (**Multi-stage**, face, human pose)
    - Background clutter removal (SDN)
    - Short and long range temporal relationship (Li fei-fei and Yu kai's works)

196

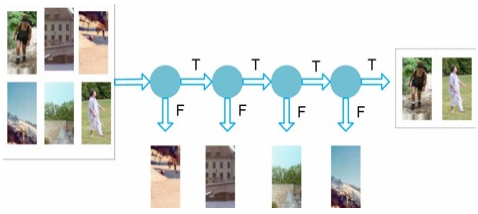
Multi-Stage Contextual Deep Learning

X. Zeng, W. Ouyang and X. Wang, "Multi-Stage Contextual Deep Learning for Pedestrian Detection," ICCV 2013

197

Motivated by Cascaded Classifiers and Contextual Boost

- The classifier of each stage deals with a specific set of samples
- The score map output by one classifier can serve as contextual information for the next classifier

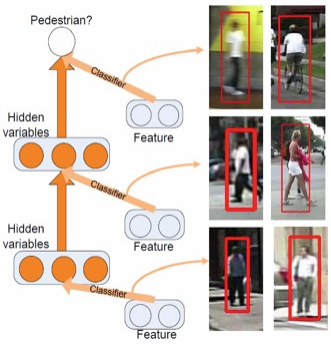


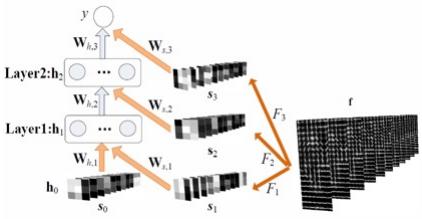
❖Only pass one detection score to the next stage
❖Classifiers are trained sequentially

Conventional cascaded classifiers for detection

198

- Simulate the cascaded classifiers by mining hard samples to train the network stage-by-stage
- Cascaded classifiers are jointly optimized instead of being trained sequentially
- The deep model keeps the score map output by the current classifier and it serves as contextual information to support the decision at the next stage
- To avoid overfitting, a stage-wise pre-training scheme is proposed to regularize optimization

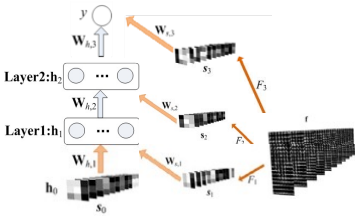




199

Training Strategies

- Unsupervised pre-train $\mathbf{W}_{h,i+1}$ layer-by-layer, setting $\mathbf{W}_{s,i+1} = 0$, $\mathbf{F}_{i+1} = 0$
- Fine-tune all the $\mathbf{W}_{h,i+1}$ with supervised BP
- Train $\mathbf{F}_{i+1}$ and $\mathbf{W}_{s,i+1}$ with BP stage-by-stage
- A correctly classified sample at the previous stage does not influence the update of parameters
- Stage-by-stage training can be considered as adding regularization constraints to parameters, i.e. some parameters are constrained to be zeros in the early training stages



Log error function:

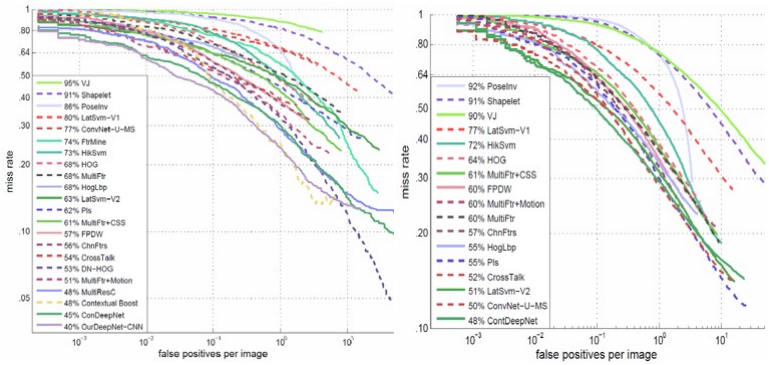
$$E = -l \log y - (1 - l) \log (1 - y)$$

Gradients for updating parameters:

$$d\theta_{i,j} = -\frac{\partial E}{\partial \theta_{i,j}} = -\frac{\partial E}{\partial y} \frac{\partial y}{\partial \theta_{i,j}} = -(y - l) \frac{\partial y}{\partial \theta_{i,j}}$$

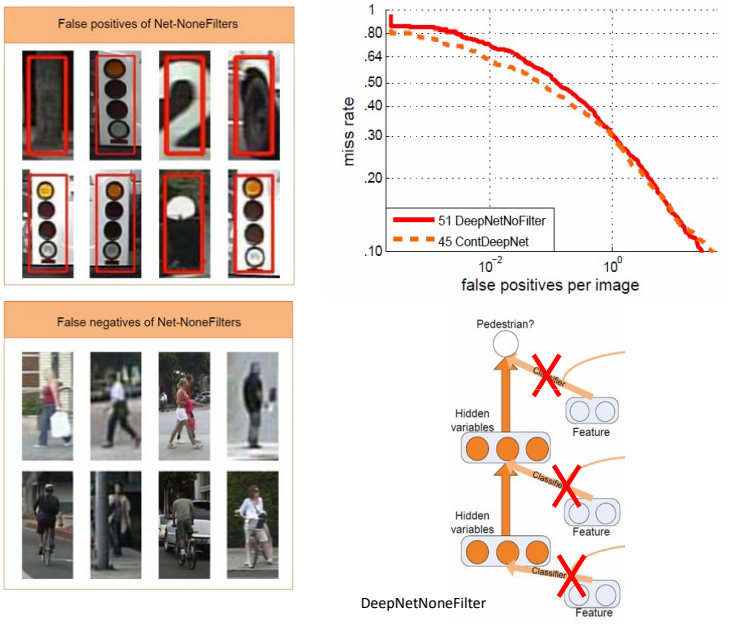
200

Experimental Results



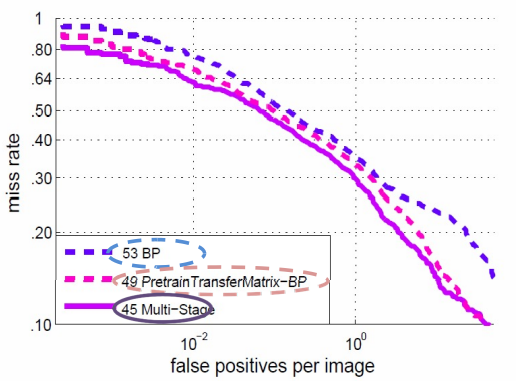
Caltech ETH Z

201



202

Comparison of Different Training Strategies



Network-BP: use back propagation to update all the parameters without pre-training
PretrainTransferMatrix-BP: the transfer matrices are unsupervised pertrained, and then all the parameters are fine-tuned
Multi-stage: our multi-stage training strategy

203

2 "How to"s

- How to effectively train a deep model
  - Data augmentation
  - Label more data
  - Pre-train on large-scale related data (RCNN)
  - Layerwise pre-training + fine tuning (Multi-stage)
- How to formulate a vision problem with deep learning
  - Tune hyper-parameters, e.g. number of hidden nodes, number of layers, activation function, dropout.
  - **Make use of experience and insights obtained in CV research**
    - Sequential design/learning vs joint learning
    - Contextual information (Multi-stage, face, human pose)
    - **Background clutter removal (SDN)**
    - Short and long range temporal relationship (Action recognition)

204

Switchable Deep Network

P. Luo, Y. Tian, X. Wang, and X. Tang, "Switchable Deep Network for Pedestrian Detection", CVPR 2014

205

Switchable Deep Network for Pedestrian Detection

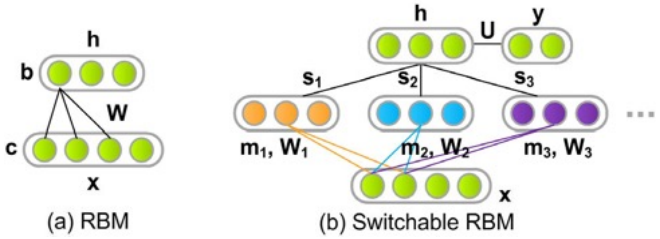
- *Background clutter* and large variations of pedestrian appearance.
- **Proposed Solution.** A Switchable Deep Network (SDN) for learning the foreground map and removing the effect background clutter.

206

Switchable Deep Network for Pedestrian Detection

- Switchable Restricted Boltzmann Machine

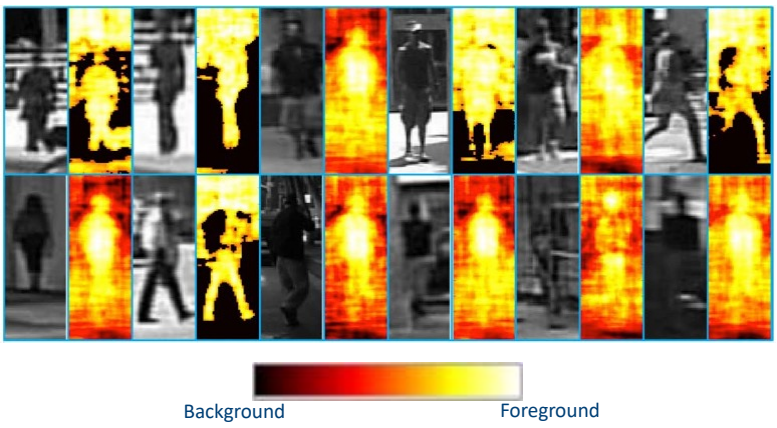
$$E(\mathbf{x},\mathbf{y},\mathbf{h},\mathbf{s},\mathbf{m};\Theta) = -\sum_{k=1}^{K} s_k \mathbf{h}_k^T (\mathbf{W}_k(\mathbf{x}\circ\mathbf{m}_k)+\mathbf{b}_k) - \sum_{k=1}^{K} s_k \mathbf{c}_k^T(\mathbf{x}\circ\mathbf{m}_k) - \mathbf{y}^T\mathbf{U}\sum_{k=1}^{K} s_k\mathbf{h}_k - \mathbf{d}^T\mathbf{y},$$



207

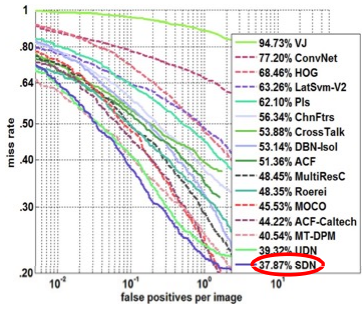
Switchable Deep Network for Pedestrian Detection

- Switchable Restricted Boltzmann Machine

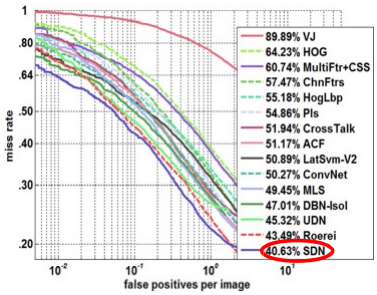


208

Switchable Deep Network for Pedestrian Detection



(a) Performance on Caltech Test



(b) Performance on ETH

209

2 "How to"s

- How to effectively train a deep model
  - Data augmentation
  - Label more data
  - **Pre-train on large-scale related data (RCNN)**
  - **Layerwise pre-training + fine tuning (Multi-stage)**

  How to formulate a vision problem with deep learning
  - Tune hyper-parameters, e.g. number of hidden nodes, number of layers, activation function, dropout.
  - **Make use of experience and insights obtained in CV research**
    - Sequential design/learning vs **joint learning**
    - Contextual information (Multi-stage, face, human pose)
    - Background clutter removal (SDN)
    - Short and long range temporal relationship (Action recognition)

210

Human part localization

- Facial Keypoint Detection
- Human pose estimation

Sun et al. CVPR' 13

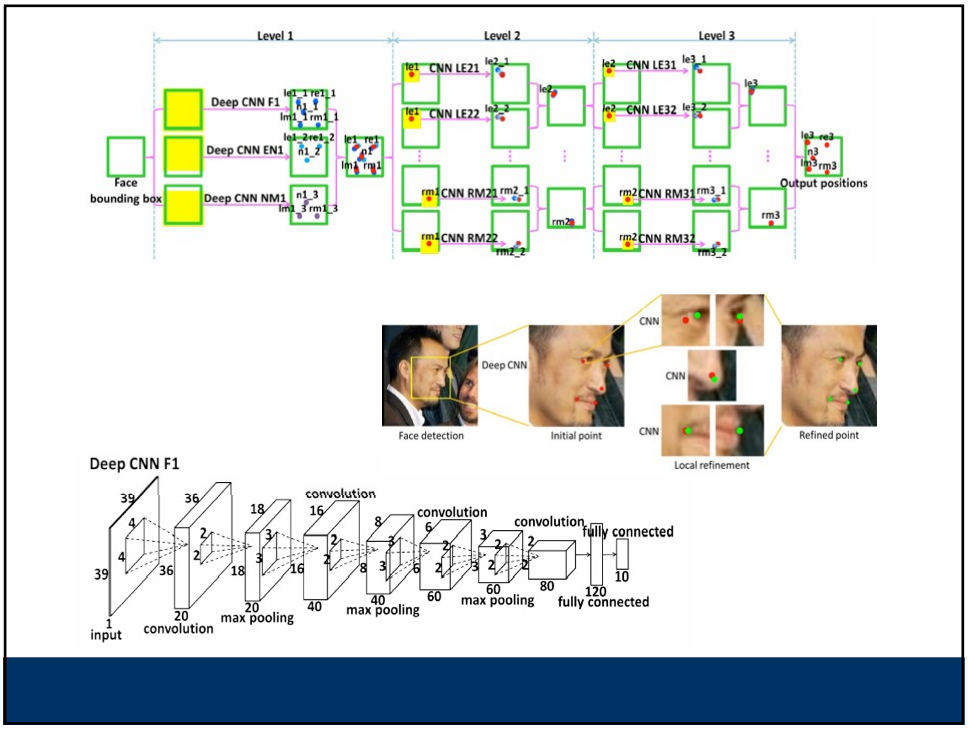
Ouyang et al. CVPR' 14

211

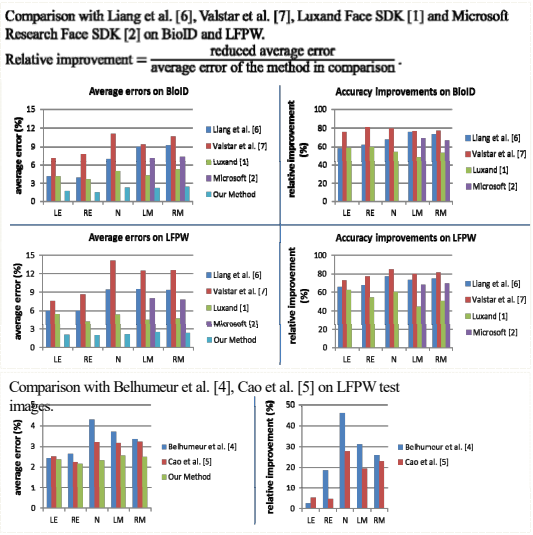
Facial Keypoint Detection

- Y. Sun, X. Wang and X. Tang, "Deep Convolutional Network Cascade for Facial Point Detection," CVPR 2013

212

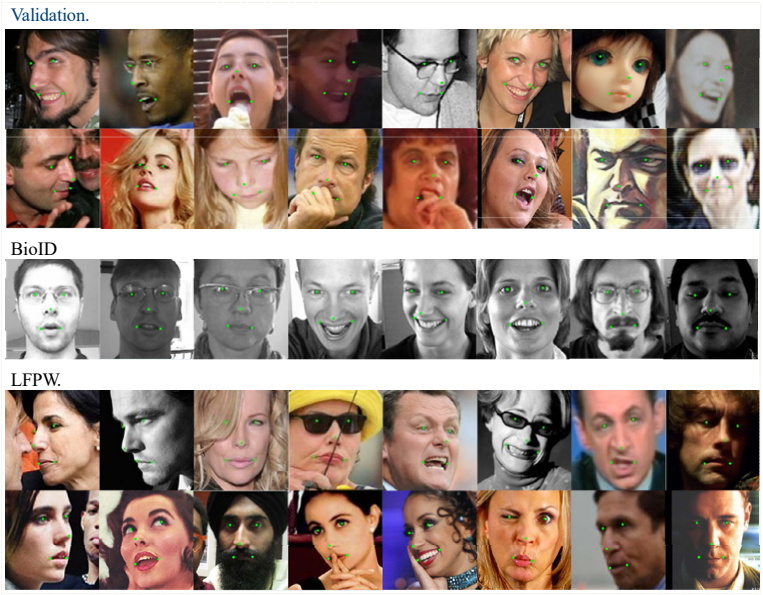


213



1. http://www.luxand.com/facesdk/, http://research.microsoft.com/en-us/projects/facesdk/.
2. Jesorsky, K. J. Kirchberg, and R. Frischholz. Robust face detection using the hausdorff distance. In Proc. AVBPA, 2001.
3. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In Proc. CVPR, 2011.
4. X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In Proc. CVPR, 2012.
5. Liang, R. Xiao, F. Wen, and J. Sun. Face alignment via component-based discriminative search. In Proc. ECCV, 2008.
6. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In Proc. CVPR, 2010.

214



215

Benefits of Using Deep Model

- The first network that takes the whole face as input needs **deep** structures to extract **high-level** features
- Take the full face as input to make full use of texture context information over the entire face to locate each keypoint
- Since the networks are trained to predict all the keypoints simultaneously, the geometric constraints among keypoints are implicitly encoded
- Global geometric constraints among keypoints can also be explicitly learned by deep model.

216

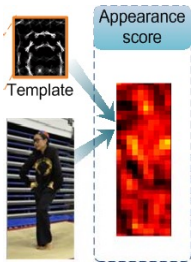
Human pose estimation

- W. Ouyang, X. Chu and X. Wang, "Multi-source Deep Learning for Human Pose Estimation" CVPR 2014.

217

Multiple information sources

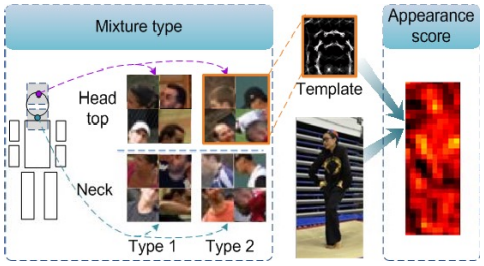
- Appearance



218

Multiple information sources

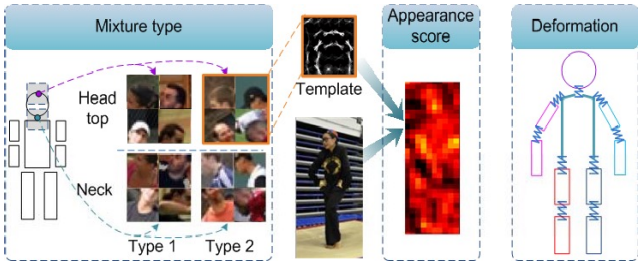
- Appearance
- Appearance mixture type



219

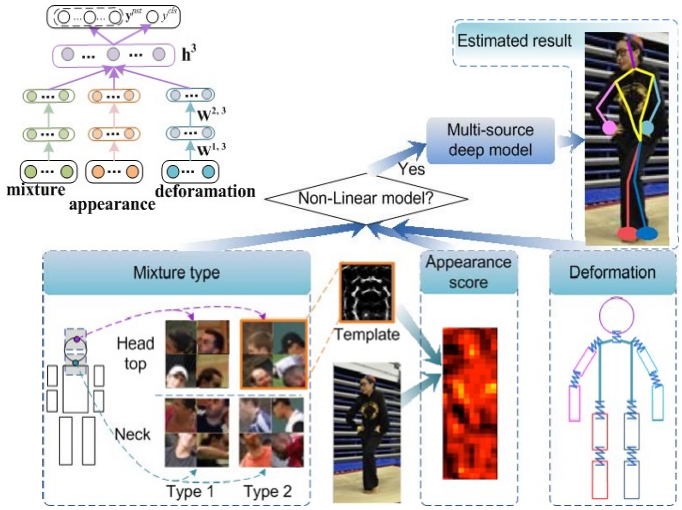
Multiple information sources

- Appearance
- Appearance mixture type
- Deformation



220

Multi-source deep model



221

Experimental results

PARSE							
Method	**Torso**	**U.leg**	**L.leg**	**U.arm**	**L.arm**	**head**	**Total**
Yang&Ramanan CVPR'11	82.9	68.8	60.5	63.4	42.4	82.4	63.6
Multi-source deep learning	89.3	78.0	72.0	67.8	47.8	89.3	71.0

UIUC People							
Method	**Torso**	**U.leg**	**L.leg**	**U.arm**	**L.arm**	**head**	**Total**
Yang&Ramanan CVPR'11	81.8	65.0	55.1	46.8	37.7	79.8	57.0
Multi-source deep learning	89.1	72.9	62.4	56.3	47.6	89.1	65.6

LSP							
Method	**Torso**	**U.leg**	**L.leg**	**U.arm**	**L.arm**	**head**	**Total**
Yang&Ramanan CVPR'11	82.9	70.3	67.0	56.0	39.8	79.3	62.8
Multi-source deep learning	85.8	76.5	72.2	63.3	46.6	83.1	68.6

Up to 8.6 percent accuracy improvement with global geometric constraints

222

Experimental results

Left: mixtire-of-parts (Yang&Ramanan CVPR'11) Right: Multi-source deep learning

223

2 "How to"s

- How to effectively train a deep model
  - Data augmentation
  - Label more data
  - **Pre-train on large-scale related data (RCNN)**
  - Layerwise pre-training + fine tuning (Multi-stage)

How to formulate a vision problem with deep learning

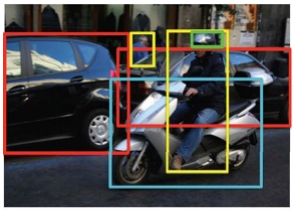
- Tune hyper-parameters, e.g. number of hidden nodes, number of layers, activation function, dropout.
- Make use of experience and insights obtained in CV research
  - Sequential design/learning vs joint learning
  - Contextual information (Multi-stage, face, human pose)
  - Background clutter removal (SDN)
  - Short and long range temporal relationship (Li fei-fei and Yu kai's works)

224

Object detection

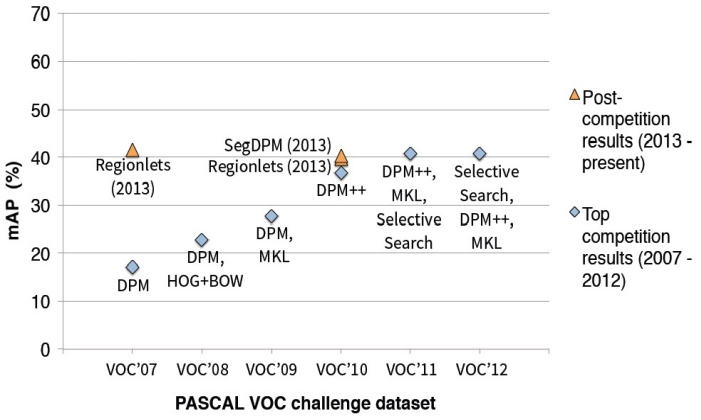
Pascal VOC
~ 20 object classes
Training: ~ 5,700 images
Testing: ~10,000 images

Image-net ILSVRC
~ 200 object classes
Training: ~ 395,000 images
Testing: ~ 40,000 images



225

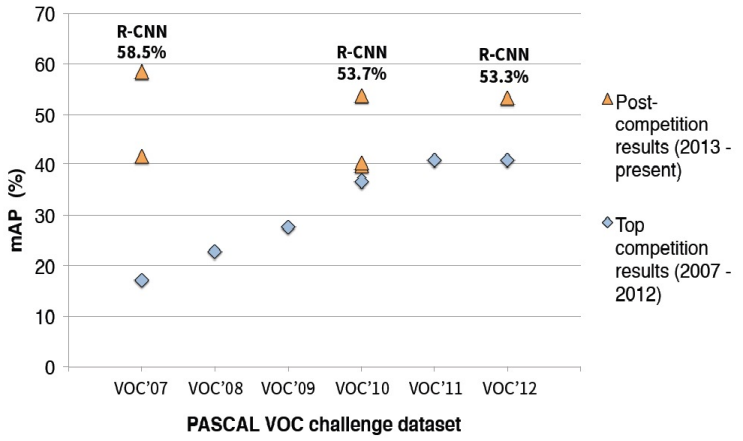
SIFT, HOG, LBP, DPM …



[Regionlets. Wang et al. ICCV'13] [SegDPM. Fidler et al. CVPR'13]

226

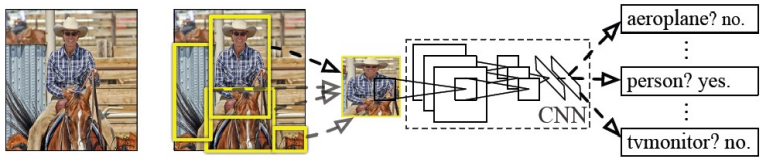
With CNN features



R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," CVPR, 2014.

227

R-CNN: regions + CNN features



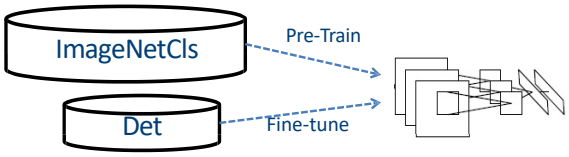
Region:
91.6%/98% recall rate on ImageNet/PASCAL
Selective Search [van de Sande, Uijlings et al. IJCV 2013].

Deep model from Krizhevsky, Sutskever & Hinton. NIPS 2012 SVM: Liblinear

228

RCNN: deep model training

- Pretrain for the 1000-way ILSVRC image classification task (1.2 million images)
- Fine-tune the CNN for detection
  - Transfer the representation learned from ILSVRC Classification to PASCAL (or ImageNet) detection



Network from Krizhevsky, Sutskever & Hinton. NIPS 2012 Also called "AlexNet"

229

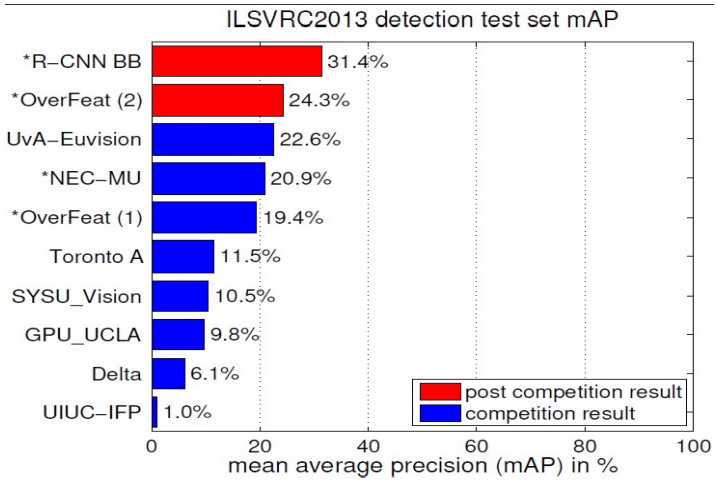
Overfeat [Sermanet et al. 2014]

- More considerations on deep model design
  - Multi-resolution, dense pooling
- Sliding window (not region proposal)
- Does not use ImageNet-Cls for pretraining

Layer	1	2	3	4	5	6	7	Output 8
Stage	conv + max	conv + max	conv	conv	conv + max	full	full	full
# channels	96	256	512	1024	1024	3072	4096	1000
Filter size	11x11	5x5	3x3	3x3	3x3	-	-	-
Conv. stride	4x4	1x1	1x1	1x1	1x1	-	-	-
Pooling size	2x2	2x2	-	-	2x2	-	-	-
Pooling stride	2x2	2x2	-	-	2x2	-	-	-
Zero-Padding size	-	-	1x1x1x1	1x1x1x1	1x1x1x1	-	-	-
Spatial input size	231x231	24x24	12x12	12x12	12x12	6x6	1x1	1x1

230

Experimental results on ILSVRC 2013



231

Experimental results on ILSVRC 2014

Rank	Name	Mean AP	Descriptio
1	GoogLeNe t	0.4393 3	Deep learning
2	CUHK	0.4065	Deep
3	DeepInsigh t	0.4045 2	Deep learning
	Uv -Euvision	0.3542 1	Deep learning
5	Berkley Vision	0.3452 1	Deep learning

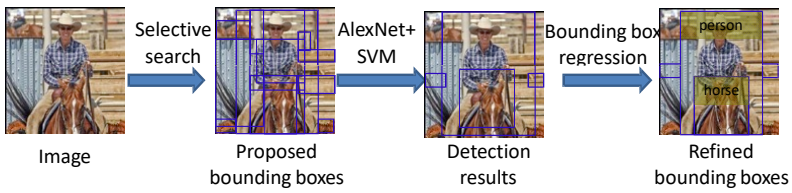
232

DeepID-Net: deformable deep convolutional neural networks for generic object detection

W. Ouyang, et al. "DeepID-Net: multi-stage and deformable deep convolutional neural networks for object detection," arXiv:1409.3505, 2014.

233

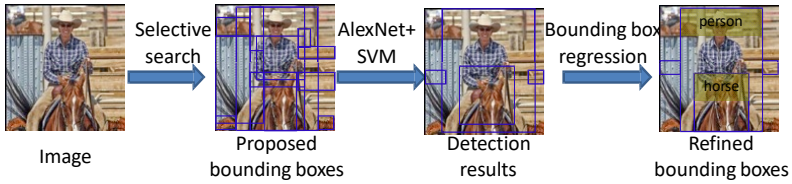
RCNN



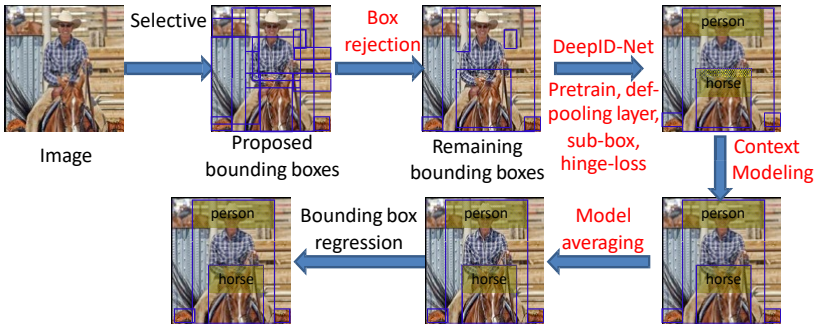
234

RCNN

Mean ap 31.4 ➡ to 40.67 (new result on)

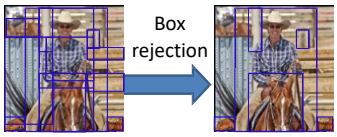


DeepID-Net



235

Bounding box rejection



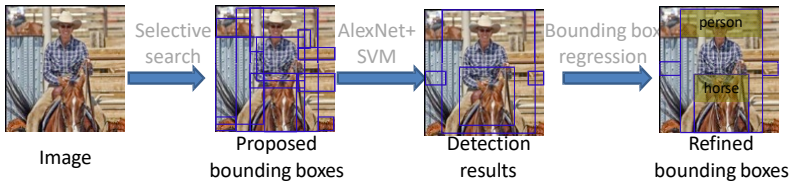
- Motivation
  - Speed up feature extraction by ~10 times
  - Improve mean AP by 1%
- RCNN
  - Selective search: ~ 2400 bounding boxes per image
  - ILSVRC val: ~20,000 images, ~2.4 days
  - ILSVRC test: ~40,000 images, ~4.7days
- Bounding box rejection by RCNN:
  - For each box, RCNN has 200 scores $S_{1...200}$ for 200 classes
  - If $\max(S_{1...200}) < -1.1$, reject. 6% remaining bounding boxes

Remaining window	100%	20%	6%
Recall (val_1)	92.2%	89.0%	84.4%
Feature extraction time (seconds per image)	10.24	2.88	1.18

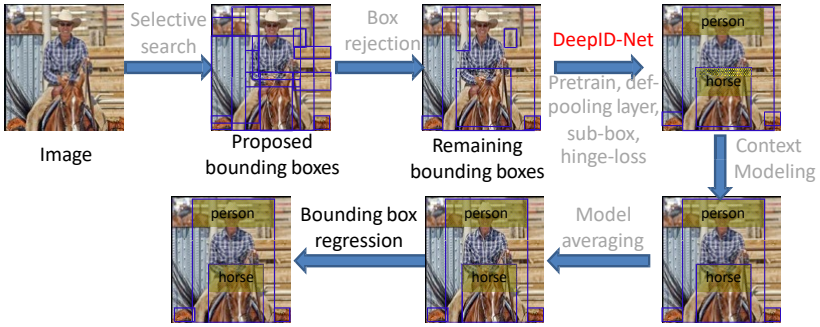
Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *CVPR*, 2014

236

RCNN

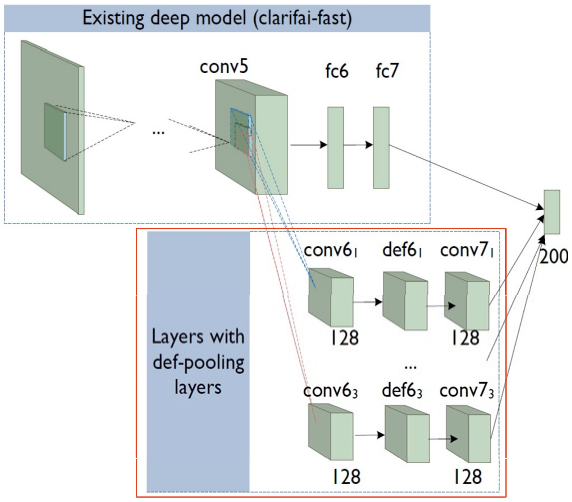


DeepID-Net



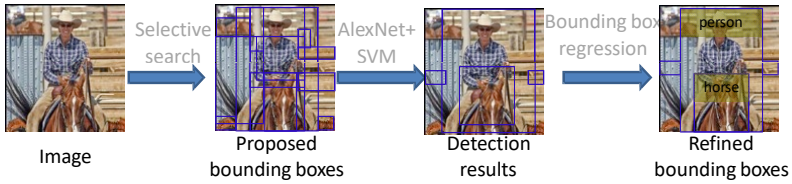
237

DeepID-Net

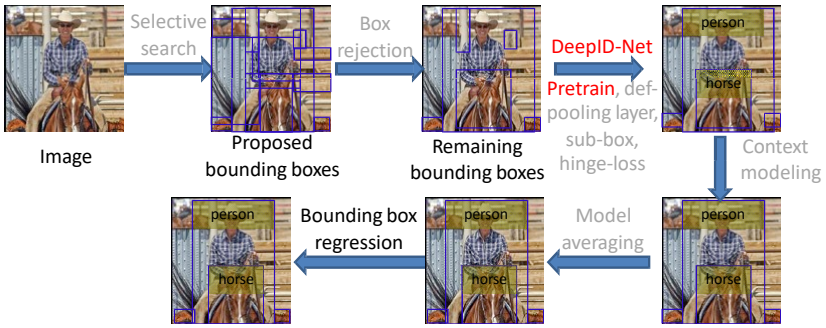


238

RCNN



DeepID-Net



239

Pretrain the deep model

- RCNN (Cls+Det)
  - AlexNet
  - Pretrain on image-level annotation data with 1000 classes
  - Finetune on object-level annotation data with 200+1 classes
- Investigation
  - Classification vs. detection (image vs. tight bounding box)?
  - 1000 classes vs. 200 classes
  - AlexNet or Clarifai or other choices, e.g. GoogleLenet?
  - Complementary

240

Deep model training – pretrain

- RCNN (Cls+Det)
  - Pretrain on image-level annotation with 1000 classes
  - Finetune on object-level annotation with 200 classes
  - Gap: classification vs. detection, 1000 vs. 200

Image classification

Object detection

241

Deep model training – pretrain

- RCNN (Cls+Det)
  - Pretrain on image-level annotation with 1000 classes
  - Finetune on object-level annotation with 200 classes
  - Gap: classification vs. detection, 1000 vs. 200
- DeepID-Net (Loc+Det)
  - Pretrain on object-level annotation with 1000 classes
  - Finetune on object-level annotation with 200 classes

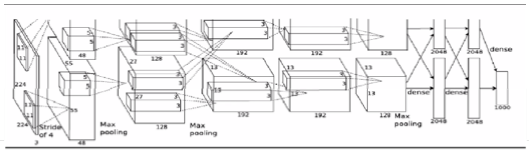
Training scheme	Cls+Det	Cls+Det	Cls+Loc+Det	Loc+Det
Net structure	AlexNet	Clarifai	Clarifai	Clarifai
mAP (%) on val2	29.9	31.8	33.4	36.0

242

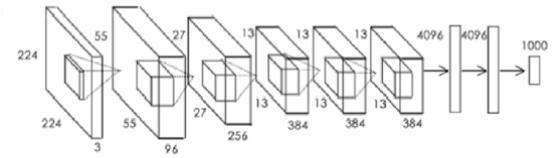
Deep model design

- AlexNet or Clarifai [Zeiler 2013]

Clarifai



AlexNet



Net structure	AlexNet	AlexNet	Clarifai
Annotation level	Image	Object	Object
Bbox rejection	n	n	n
mAP (%)	29.9	34.3	35.6

243

Deep model training – pretrain

- RCNN (ImageNet Cls+Det)
  - Pretrain on image-level annotation with 1000 classes
  - Finetune on object-level annotation with 200 classes
  - Gap: classification vs. detection, 1000 vs. 200
- DeepID-Net (ImageNet Cls+Loc+Det)
  - Pretrain on image-level annotation with 1000 classes
  - Finetune on object-level annotation with 1000 classes
  - Finetune on object-level annotation with 200 classes

Training scheme	Cls+Det	Cls+Det	Cls+Loc+Det
Net structure	AlexNet	Clarifai	Clarifai
mAP (%) on val2	29.9	31.8	33.4

244

Result and discussion

- RCNN (Cls+Det),
- Investigation
  - Better pretraining on 1000 classes

	Image annotation
200 classes (Det)	20.7
1000 classes (Cls-Loc)	31.8

245

Result and discussion

- RCNN (Cls+Det),
- Investigation
  - Better pretraining on 1000 classes
  - Object-level annotation is more suitable for pretraining

	Image annotation	Object annotation
200 classes (Det)	20.7	28.0
1000 classes (Cls-Loc)	31.8	36

23% AP increase for rugby ball

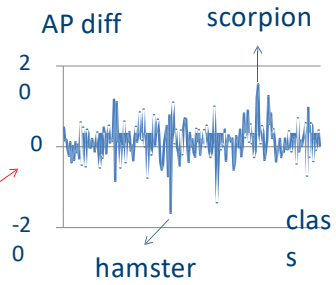
17.4% AP increase for hammer

246

Result and discussion

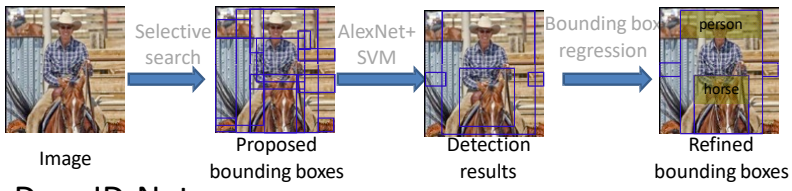
- RCNN (ImageNet Cls+Det),
- Investigation
  - Better pretraining on 1000 classes
  - Object-level annotation is more suitable for pretraining
  - Clarifai is better. But Alex and Clarifai are complementary on different classes.

Net structure	AlexNet	AlexNet	Clarifai
Annotation level	Image	Object	Object
Bbox rejection	n	n	n
mAP（%）	29.9	34.3	35.6

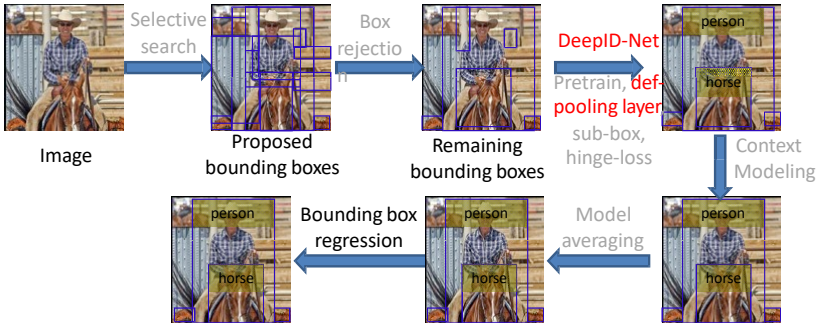


247

RCNN



DeepID-Net



248

Deep model training – def-pooling layer

- RCNN (ImageNet Cls+Det)
  - Pretrain on image-level annotation with 1000 classes
  - Finetune on object-level annotation with 200 classes
  - Gap: classification vs. detection, 1000 vs. 200
- DeepID-Net(ImageNet Loc+Det)
  - Pretrain on object-level annotation with 1000 classes
  - Finetune on object-level annotation with 200 classes **with def-pooling** layers

Net structure	Without Def Layer	With Def layer
mAP (%) on val2	36.0	38.5

249

Deformation

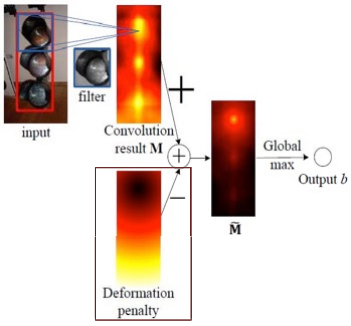
- Learning deformation [a] is effective in computer vision society.
- Missing in deep model.
- We propose a new deformation constrained pooling layer.

[a] P. Felzenszwalb, R. B. Grishick, D.McAllister, and D. Ramanan. Object detection with discriminatively trained part based models. IEEE Trans. PAMI, 32:1627–1645, 2010.

250

Deformation Layer [b]

$$\mathbf{B}_p = \mathbf{M}_p + \sum_{n=1}^{N} c_{n,p}\mathbf{D}_{n,p} \qquad s_p = \max_{(x,y)} b_p^{(x,y)}$$

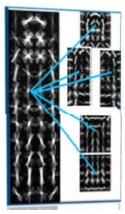


[b] Wanli Ouyang, Xiaogang Wang, "Joint Deep Learning for Pedestrian Detection", ICCV 2013.

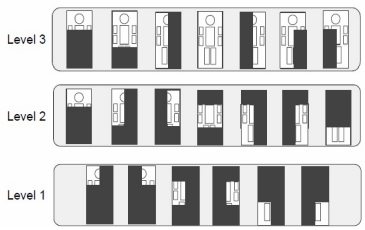
251

Modeling Part

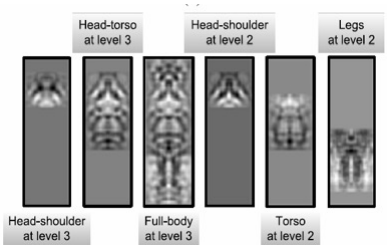
- Different parts have different sizes
- Design the filters with variable sizes



Part models learned from HOG



Part models



Learned filtered at the second convolutional layer

252

Deformation layer for repeated patterns

Pedestrian	General object
Assume no repeated pattern	Repeated patterns

253

Deformation layer for repeated patterns

Pedestrian detection	General object detection
Assume no repeated pattern	Repeated patterns
Only consider one object class	Patterns shared across different object classes





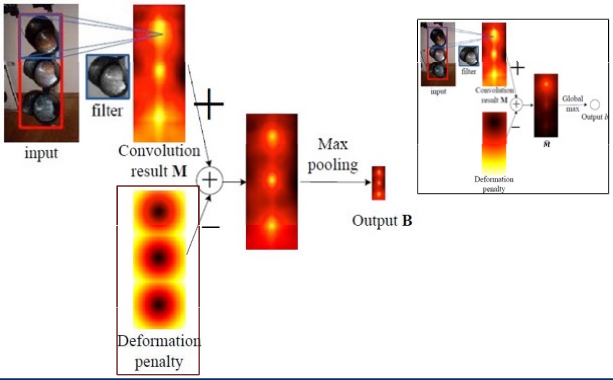



254

Deformation constrained pooling layer

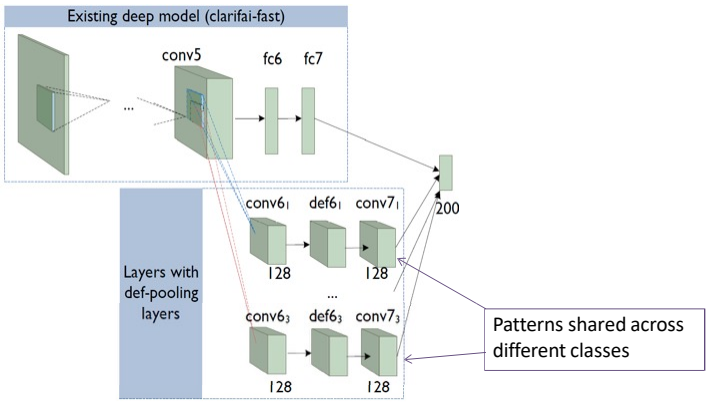
Can capture multiple patterns simultaneously

$$b^{(x,y)} = \max_{i,j \in \{-R,\cdots,R\}} \{m^{(k_x \cdot x+i, k_y \cdot y+j)} - \sum_{n=1}^{N} c_n d_n^{i,j}\},$$



255

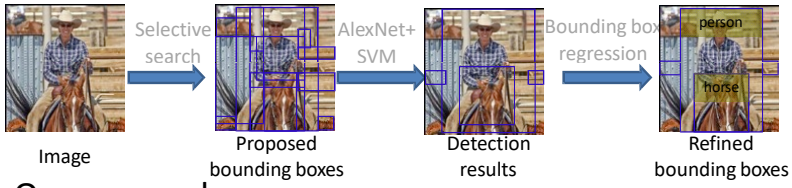
A deep model with deformation layer



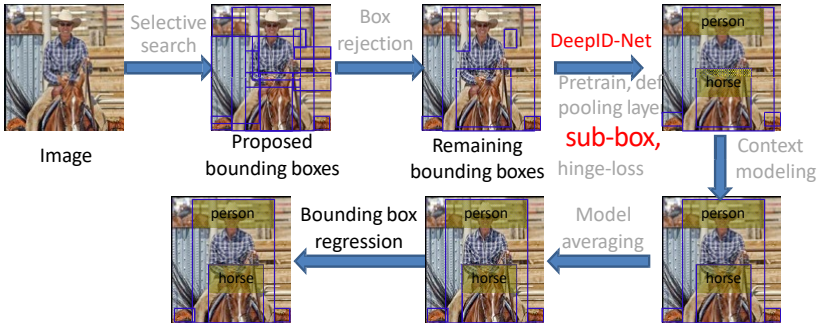
Training scheme	Cls+Det	Loc+Det	Loc+Det
Net structure	AlexNet	Clarifai	Clarifai+Def layer
Mean AP on val2	0.299	0.360	0.385

256

RCNN



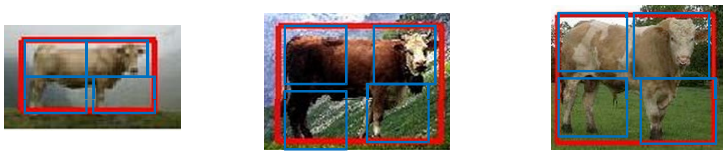
Our approach



257

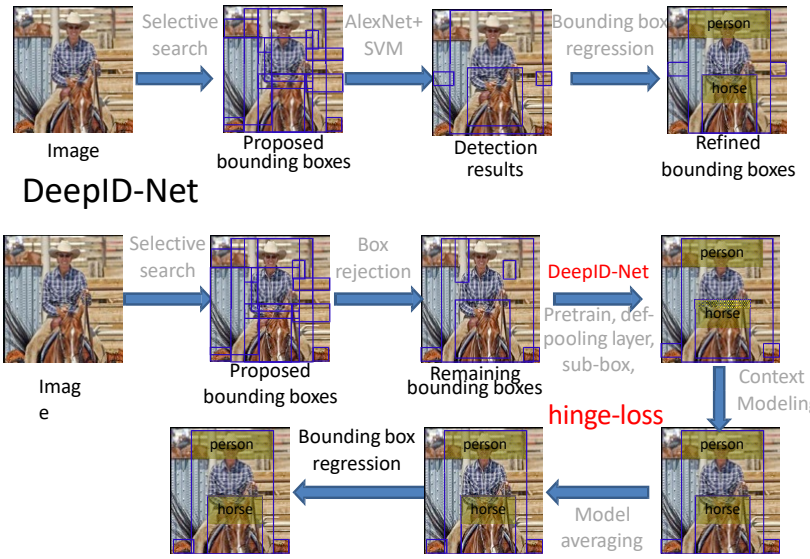
Sub-box features

- Take the per-channel max/average features of the last fully connected layer from 4 subboxes of the root window.
- Concatenate subbox features and the features in the root window. Learn an SVM for combining these features.
- Subboxes are proposed regions that has >0.5 overlap with the four quarter regions. Need not compute features.
- 0.5 mAP improvement.
- So far not combined with deformation layer. Used as one of the models in model averaging



258

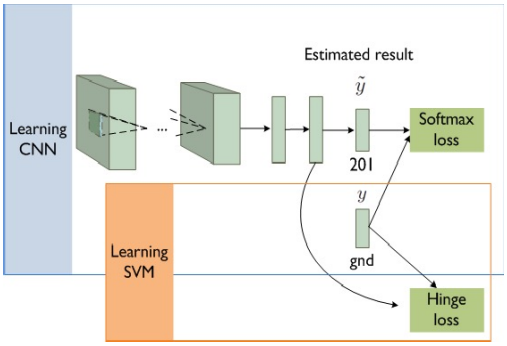
RCNN



259

Deep model training – SVM-net

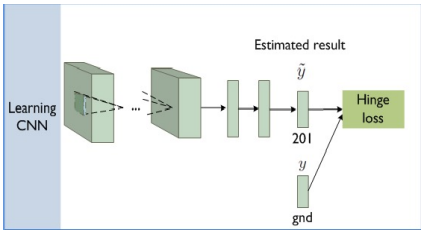
- RCNN
  - Fine-tune using soft-max loss (Softmax-Net)
  - Train SVM based on the fc7 features of the fine-tuned net.



260

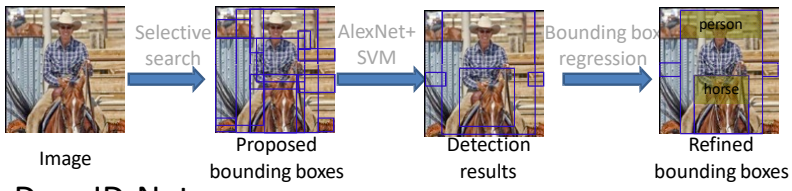
Deep model training – SVM-net

- RCNN
  - Fine-tune using soft-max loss (Softmax-Net)
  - Train SVM based on the fc7 features of the fine-tuned net.
- Replace Soft-max loss by Hinge loss when fine-tuning (SVM-Net)
  - Merge the two steps of RCNN into one
  - Require no feature extraction from training data (~60 hours)

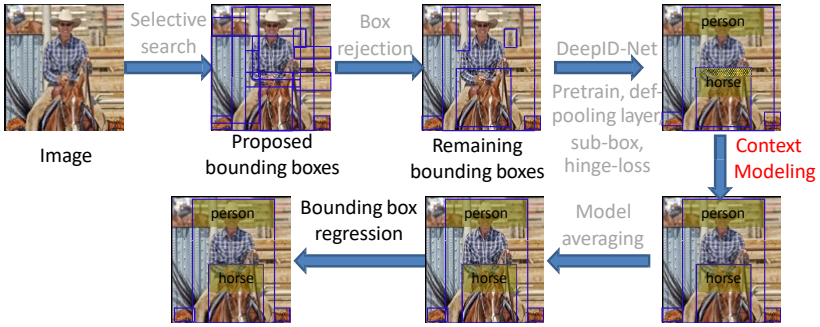


261

RCNN



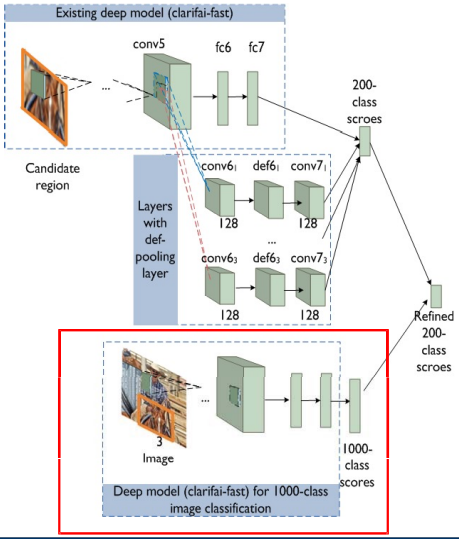
DeepID-Net



262

Context modeling

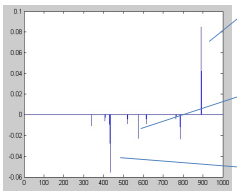
- Use the 1000 class Image classification score.
- ~1% mAP improvement.

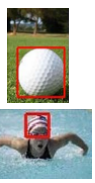


263

Context modeling

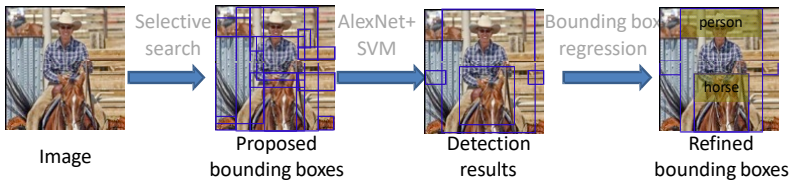
- Use the 1000-class Image classification score.
  - ~1% mAP improvement.
  - Volleyball: improve ap by 8.4% on val2.



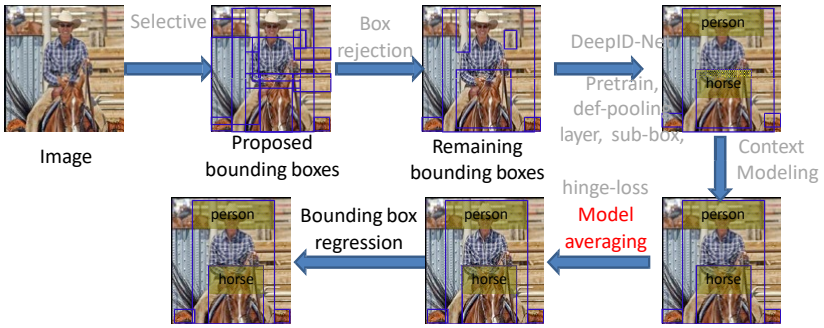


264

RCNN



DeepID-Net



265

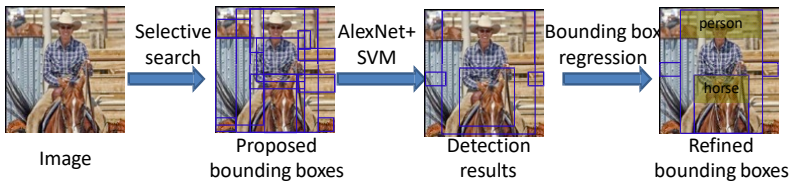
Model averaging

- Not only change parameters
  - Net structure: AlexNet(A), Clarifai (C), Deep-ID Net (D), DeepID Net2 (D2)
  - Pretrain: Classification (C), Localization
  - Region rejection or not
  - Loss of net, softmax (S), Hinge loss (H)
  - Choose different sets of models for different object class

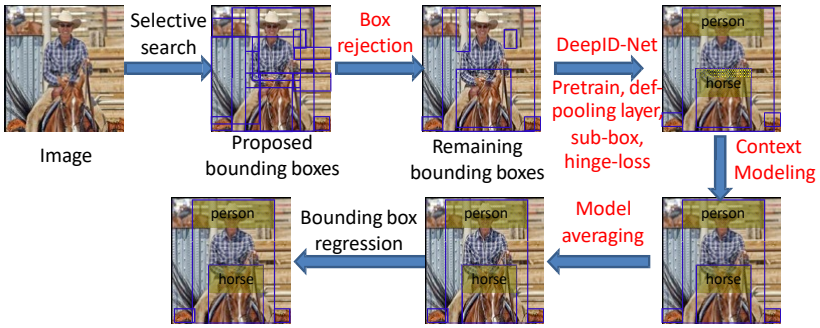
Model	1	2	3	4	5	6	7	8	9	10
Net structure	A	A	C	C	D	D	D2	D	D	D
Pretrain	C	C+L	C	C+L	C+L	C+L	L	L	L	L
Reject region?	Y	N	Y	Y	Y	Y	Y	Y	Y	Y
Loss of net	S	S	S	H	H	H	H	H	H	H
Mean ap	0.31	0.312	0.321	0.336	0.353	0.36	0.37	0.37	0.371	0.374

266

RCNN



DeepID-Net

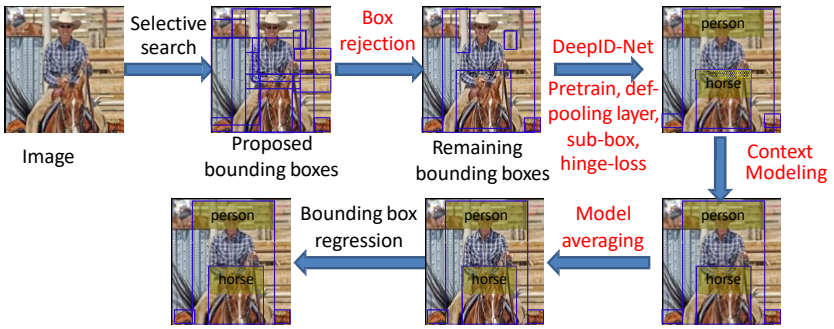


267

Component analysis

Detection Pipeline	RCNN	Box rejection	Clarifai	Loc+ Det	+Def layer	+cont ext	+bbox regr.	Model avg.
mAP on val2	29.9	30.9	31.8	36.0	38.5	39.2	40.1	42.4
mAP on test					38.0	38.6	39.4	41.7

DeepID-Net



268

Summary

- 1. Bounding rejection. Save feature extraction by about 10 times, slightly improve mAP (~1%).
- 2. Pre-training with object-level annotation, more classes. 4.2% mAP
- 3. Def-pooling layer. 2.5% mAP
- 4. Hinge loss. Save feature computation time (~60 h).
- 5. Model averaging. Different model designs and training schemes lead to high diversity

269

Conclusion - 2 "How to's"

- How to effectively train a deep model
  - Data augmentation
  - Label more data
  - **Pre-train on large-scale related data (RCNN)**
  - **Layerwise pre-training + fine tuning (Multi-stage)**
- How to formulate a vision problem with deep learning?
  - Tune hyper-parameters, e.g. number of hidden nodes, number of layers, activation function, dropout.
  - **Make use of experience and insights obtained in CV research**
    - Sequential design/learning vs **joint learning**
    - Contextual information (Multi-stage, face, human pose)
    - Background clutter removal (SDN)
    - Short and long range temporal relationship (Action recognition)

270

Outline

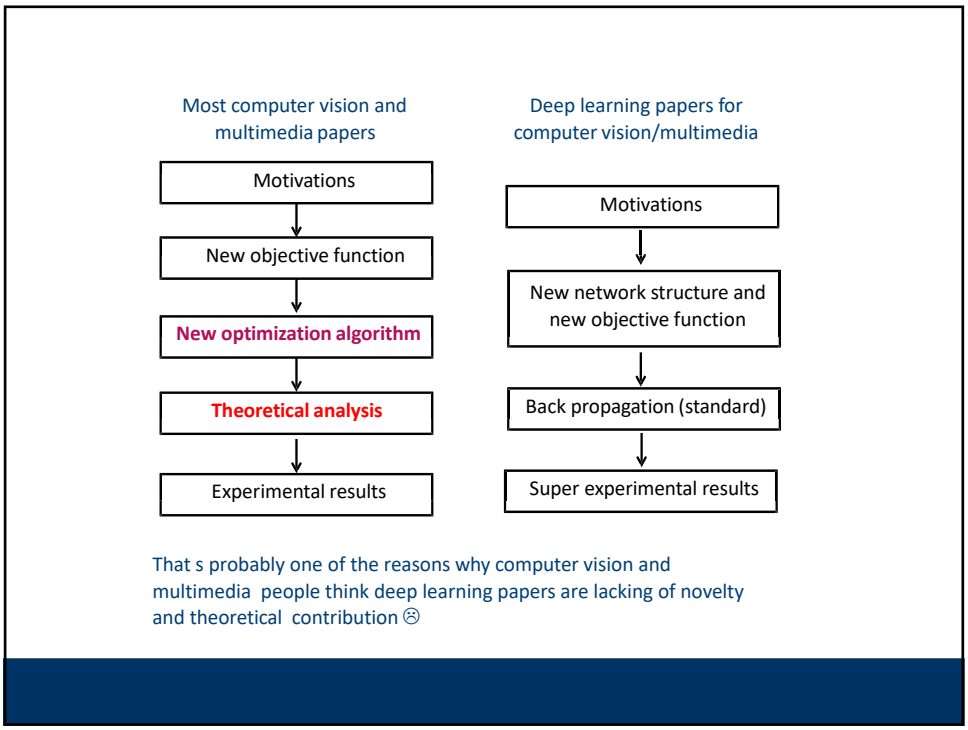
- Introduction to deep learning
- Deep learning for object recognition
- Deep learning for object segmentation
- Deep learning for object detection
- Open questions and future works

273

« Concerns » about deep learning

- C1: Weak on theoretical support (convergence, bound, local minimum, why it works)
  - **It's true.** That's why deep learning papers were not accepted by the computer vision/multimedia community for a long time. **Any theoretical studies in the future are important.**

274



That s probably one of the reasons why computer vision and multimedia people think deep learning papers are lacking of novelty and theoretical contribution ☹

275

« Concerns » about deep learning

- C2: It is hard for computer vision/multimedia people to have innovative contributions to deep learning. Our job becomes preparing the data + using deep learning as a black box. That's the end of research … (?)
  - **That's not true.** Computer vision and multimedia researchers have developed many systems with deep architectures. But we just didn't know how to jointly learn all the components. Our research experience and insights can help to design new deep models and pre-training strategies.
  - Many machine learning models and algorithms were motivated by computer vision and multimedia applications. However, computer vision and multimedia did not have close interaction with neural networks in the past 15 years. We expect fast development of deep learning driven by applications.

276

« Concerns » about deep learning

- C3: Since the goal of neural networks is to solve the general learning problem, why do we need domain knowledge?
  - The most successful deep model on image and video related applications is convolutional neural network, which has used domain knowledge (filtering, pooling)
  - Domain knowledge is important especially when the training data is not large enough

277

« Concerns » about deep learning

- C4: Good results achieved by deep learning come from manually tuning network structures and learning rates, and trying different initializations
  - **That's not true.** One round evaluation may take several weeks. There is no time to test all the settings.
  - Designing and training deep models does require a lot of empirical experience and insights. There are also a lot of tricks and guidance provided by deep learning researchers. Most of them make sense intuitively but without strict proof.

278

« Concerns » about deep learning

- C5: Deep learning is more suitable for industry rather than research groups in universities
  - Industry has big data and computation resources
  - Research groups from universities can contribute on model design, training algorithms and new applications

279

« Concerns » about deep learning

- C6: Deep learning has different behaviors when the scale of training data is different
  - Pre-training is useful when the training data small, but does not make big difference when the training data is large enough
  - So far, the performance of deep learning keep increasing with the size of training data. We don't see its limit yet.
  - Shall we spend more effort on data annotation or model design?

280

Future Works in DL …

- Explore deep learning in new applications
- Worthy to try if the applications require features or learning, and have enough training data
- We once had many doubts on deep. (Does it work for vision? Does it work for segmentation? Does it work for low-level vision?) But deep learning has given a lot of surprises …
- Applications will inspire many new deep models
- Incorporate domain knowledge into deep learning
- Integrate existing machine learning models with deep learning

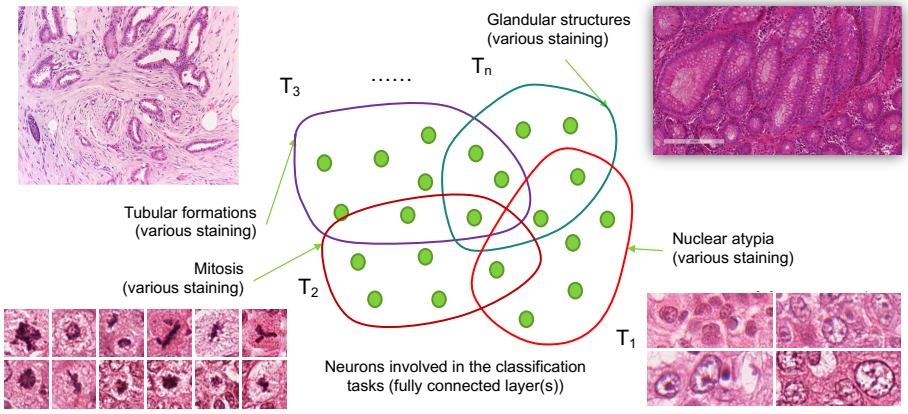
281

Future Works in DL …

- Deep learning to extract dynamic features for video analysis
- Deep models for structured data
- Theoretical studies on deep learning
- Quantitative analysis on how to design network structures and how to choose nonlinear operations of different layers in order to achieve feature invariance
- New optimization and training algorithms
- Parallel computing systems to train very large networks with larger training data

282

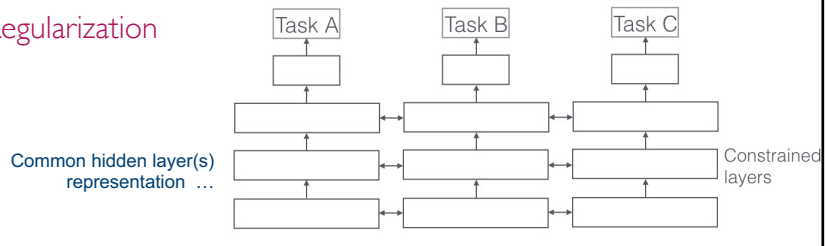
Multi Task Deep Learning



283

Why does MTL work ?

- Implicit data augmentation
- Attention focusing
- Eavesdropping (spying)
- Representation bias
- Regularization

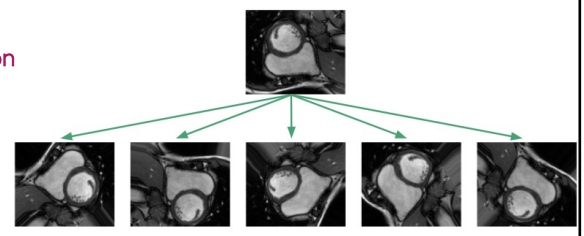


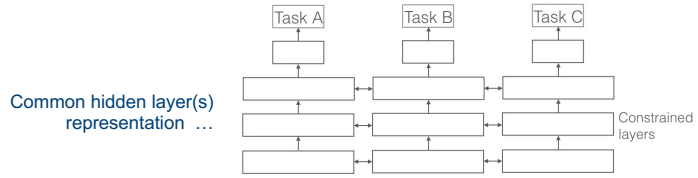
Sebastian Ruder : http://ruder.io/multi-task/

284

Why does MTL work ?

- Implicit data augmentation
- Attention focusing
- Eavesdropping (spying)
- Representation bias
- Regularization



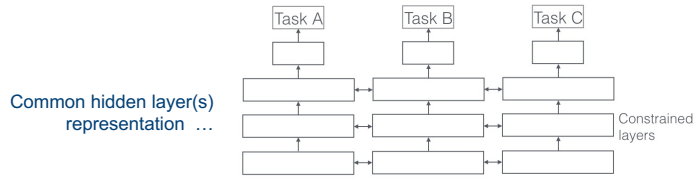


Sebastian Ruder : http://ruder.io/multi-task/

285

Why does MTL work ?

- Implicit data augmentation
- Attention focusing
- Eavesdropping (spying)
- Representation bias
- Regularization

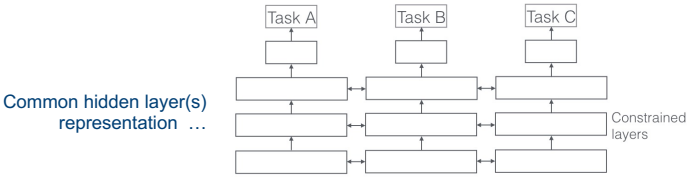


Sebastian Ruder : http://ruder.io/multi-task/

286

Why does MTL work ?

- Implicit data augmentation
- Attention focusing
- **Eavesdropping (spying)**
- Representation bias
- Regularization

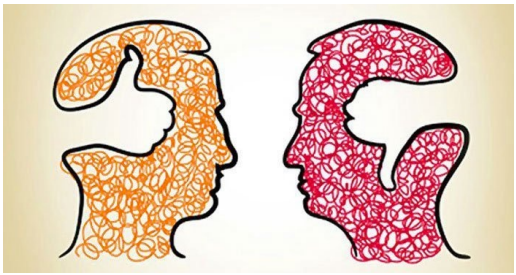


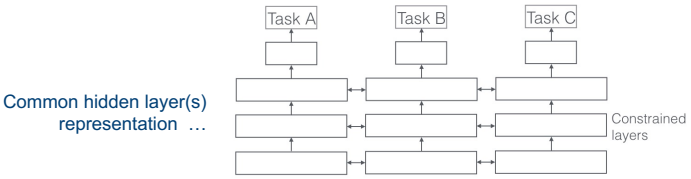
Sebastian Ruder : http://ruder.io/multi-task/

287

Why does MTL work ?

- Implicit data augmentation
- Attention focusing
- Eavesdropping (spying)
- **Representation bias**
- Regularization



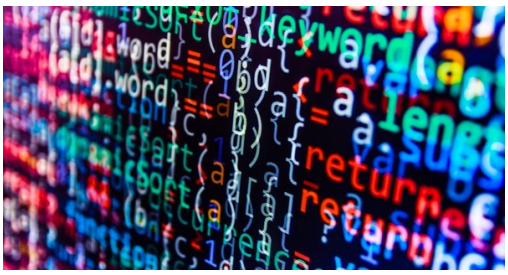


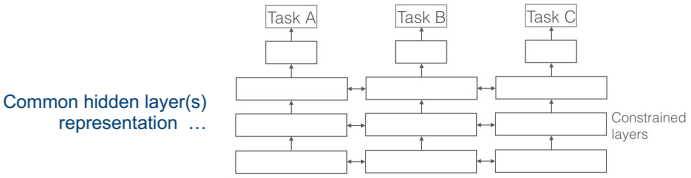
Sebastian Ruder : http://ruder.io/multi-task/

288

Why does MTL work ?

- Implicit data augmentation
- Attention focusing
- Eavesdropping (spying)
- Representation bias
- Regularization





Sebastian Ruder : http://ruder.io/multi-task/

289